COMMUNICATION BEYOND WORDS: MULTIMEDIA APPROACHES TO BRIDGING LANGUAGE DISABILITIES AND BARRIERS

Xiaojuan Ma

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

COMPUTER SCIENCE

Adviser: Perry R. Cook

September, 2010

Abstract

Human communication takes place in both spoken and written modes, as in face-to-face conversations, books reading, television watching and Internet browsing. However, for people with language disabilities (e.g. aphasia), people with low literacy, and people with poor command of a language, receiving and expressing information via a language is difficult. Because of the inability to comprehend words and / or to find words that express intended concepts, people with language barriers may encounter great challenges performing everyday tasks such as ordering food and visiting a doctor. As an alternative to words, pictures and sounds have been designed, tested, and used to evoke concepts in computer-based interfaces, education, industry, and advertisement. However, icons created by artists, user-uploaded photos, auditory icons and earcons cannot always satisfy the need for communicating everyday concepts for people with language barriers. Furthermore, the iconic vocabularies in current communication support systems have ambiguous representations and difficult to use vocabulary organization.

My thesis research addresses issues in the above paragraph and explores building multimedia support for information delivery. The work is unique in that it provides different channels (visual and auditory) to support nonverbal communication, and that it has a disambiguated sense-level network structure with semantic links between related concepts. The development of such a multimedia-enhanced lexical database is challenging, due to the lack of guidance in the creation and evaluation of multimodal representations. To construct the lexical database, I went through a cycle of design, construction, evaluation, and modification for each media modality-to-concept association. I have introduced novel representations that are under investigated (e.g. videos and environmental sounds). I have explored difference sources of multimedia data (designed by specialists, designed by general users, and web resources not tailored for communication purposes). I have experimented with new methodologies for generating and evaluating multimedia-concept associations. I have also come up with essential guidelines for better multimedia vocabulary design.

A communication support system, the Online Multimedia Language Assistant (OMLA), which employs various multimedia forms, including web images, icons, animations, videos, and environmental sounds, was built on top of my research on multimedia representations with the goal of assisting comprehension of common concepts across language barriers. OMLA is implemented as a popup dictionary in the form of a web browser extension. Users can select an unfamiliar word on a webpage to view its associated visual/auditory representation in a popup box. The Language Assistant can enhance concept understanding as people browse information on the Internet, and support face-to-face communication when people want to illustrate a term via a picture, a video or a sound.

Through a series of studies, I verified that web images are as effective as stylized icons in conveying nouns, that videos outperformed other stimuli in illustrating verbs, and that nonspeech sounds are better in distinguishing many concepts like thunder, alarms, and sneezing. The Multimedia Augmented Online Language Assistant was shown to enhance information comprehension, and I am exploring its application in real life scenarios such as medical care.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor, Professor Perry Cook, for his guidance and continuous support throughout my Ph.D. study. His vast knowledge and skills in different areas (e.g. computer music and sound, human-computer interaction, etc.), his advice and assistance on research design, conduction, and write-up/presentation, as well as his patience and understanding considerably enrich my graduate research experiences.

My hearty thanks also goes to the readers in my thesis committee: Dr. Christiane Fellbaum and Dr. Marilyn Tremaine (Rutgers University), for their immense knowledge and for their guidance and encouragement on work and life. I also want to thank the rest of the committee: Professor Szymon Rusinkiewicz and Professor Kenneth Steiglitz, for their insightful comments on my thesis research.

I must also acknowledge Professor Joanna McGrenere (Computer Science, University of British Columbia, BC, Canada), Professor Peter Graf (Psychology, UBC), Dr. Barbara Purves (Speech Language Pathologist, UBC), and Dr. Peter Johnsen (M.D., director and physician, McCosh University Health Services, Princeton University), for their generous and helpful assistance and feedback on the research projects presented in this thesis.

I owe a special thanks to Dr. Maria Klawe, who introduced me to the Princeton Aphasia Project.

This thesis would have remained impossible without my wonderful colleagues. I have received extensive help from my fellow group mates and co-authors from the Princeton Aphasia Project Sonya Nikolova, Jordan Boyd-Graber Ying, who assisted me in the design, implementation, conduction, and analysis of various experiments. I also share the credit of my work with the dear Princeton Sound Lab members: Ge Wang (the leading actor of the verb videos), Rebecca Fiebrink, Ananya Misra, and Matt Hoffman, for their great help in the design of study materials, participation in the studies and useful feedback. It was also a pleasure to work with Rock Leung (Ph.D. student in UBC) and two Princeton undergraduate students Zennen Clifton and Brett Lullo.

I would like to thank all of my participants and supporters from the Lingraphicare Inc., Adler Aphasia Center, Prince Senior Resource Center, and McCosh University Health Service Center.

I am indebted to my beloved parents, Rongyong Ma and Yongjian Li, for their constant support and care. Also I thank my dear friends in the U.S. and in China, and the fantastic staffs in the Computer Science Department, especially Ms. Melissa Lawson, for all of their help.

Lastly, I am grateful to the funding support from the National Science Foundation (under grants 0101247, 0509447 and 9984087), and the Kimberly and Frank H. Moss'71 Research Innovation Fund (Princeton School of Engineering and Applied Sciences).

Table of Contents

CHAPER 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. Natural languages including speech and/or texts, as the principal way people communicate, are an essential part of life. Daily communication not only enriches peoples' lives but also expands their vocabulary and enhances their language skills. However, for people with language disabilities, speakers of minority languages in a setting where another language dominates, and language learners, the linguistic channel of communication may be less effective. To compensate for the loss/lack of language, Augmentative and Alternative Communication (AAC) technologies introduce non-verbal representations as a replacement or supplement of words in communication. Many high-tech AAC systems have been designed for people with language impairment. However, the usage of those devices in practice is shown to be limited due to the pitfalls in the vocabulary and platform employed by the AAC systems. Little work has been done to 1) verify the communicativeness of the pictorial representations in the AAC vocabulary; 2) explore new communication media that are effective, inexpensive, and scalable; 3) improve the vocabulary structure for easy access; 4) experiment with new system platforms having fewer physical constraints.

My thesis research aims to address these questions by providing guidelines for design and methodologies to evaluate various multimedia representations (including images, icons, animations, videos, and environmental sounds) as concept carriers when words are not enough, or the right word is not known or available. An Online Multimedia Language Assistant system is designed and tested as an application of multimedia-augmented communication support bridging language disabilities and language barriers.

Natural languages have evolved to be the primary means of communication, and thus language disabilities, which are affecting a large and increasing population, greatly impact people's quality of life. For example, aphasia, one of the most common language disorders, affects approximately one million people in North America alone. Aphasia can be acquired through a stroke, brain tumor or head injuries (North America Aphasia Association2009). For individuals with aphasia, who are afflicted to varying degrees in their ability to produce or comprehend verbal or written languages, 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 (Davidson et al.2008) (Kauhanen et al.2000) (Code et al.1999). AAC techniques haven been shown to be able to help people with language disabilities to reengage in different aspects of their everyday life (Flippo et al.1995).

In a broader scope, people with unimpaired but insufficient language skills encounter similar difficulties in carrying out effective communication. This includes people who travel for fun or work to a foreign country that speaks a different language, people who study abroad or conduct international business/collaborations, people who immigrate to another country as an adult, people who are learning their first, second, or additional, and people with low literacy. Language barriers exist not only in face-to-face conversations, but also in information exchange through text, such as reading Internet posts written in an unfamiliar language. It is not an exaggeration to say that every one of us is language disabled at one point or another. One may argue that dictionaries and multilingual translation can provide enough support in the presence of language barriers. However, there are many situations where dictionaries and translation facilities lose their power, for example, there is no such term in the other language, translation devices fail to provide the correct interpretation, the explanation in the dictionary is not informative enough (too simple or with more unknown words), or the person does not know the term in his/her own language. In these cases, people facing language barriers can benefit more from Augmentative and Alternative Communication via pictures, videos, and sounds.

Most of the existing AAC devices are designed to help people with language disabilities compose sentences for communication by selecting words from a pictorial vocabulary (e.g. Lingraphica (Lingraphica2010) and DynaVox (DynaVox2010)). Little has been done on facilitating information comprehension with non-verbal representations. People still struggle with ordering food from a menu when travelling to a foreign country, and still flip through dictionaries or wrestle with Google Translate when browsing information online. Even with the AAC devices and software out in the market, user surveys and interviews with domain experts show that very few people actually use the systems to assist their daily communication. This makes AAC seem more conceptual than practical.

However, taking a deeper look at the current AAC technologies, that people are not taking advantage of them is less surprising. The effectiveness of AAC systems is greatly harmed by the inherent problems in their information encoding scheme, representation validity, vocabulary organization, and system platform.

First of all, the single picture to word mapping applied in most of the AAC vocabularies introduces ambiguities. Polysemy (same form with multiple meanings) is a common phenomenon among words, and thus assigning representations to words instead of specific meanings can miss out a lot of information. Furthermore, if the representation with the wrong meaning is used in the communication, misunderstanding may occur (Figure 1-1). People have found the current AAC vocabularies and systems misleading and insufficient to fulfill their communication needs.

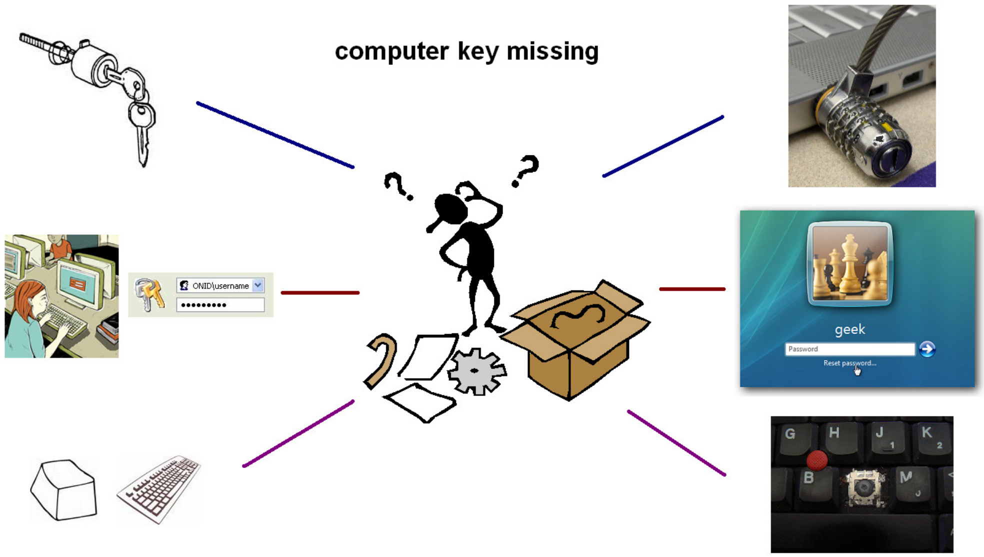


Figure 1-1. Example of ambiguous non-verbal representations.

Second, the representations provided by the current AAC systems are mostly icons (graphical drawings/symbols) with some user-uploaded photos. The existing symbol systems are all created by designers or artists, which only reflect the opinions on how information should be depicted of a small subset of people who are usually not the ones using the device. Different designers have different ideas (e.g. Figure 1-2, representations for the phrase "good or bad"). Little is known about the actual evocativeness of these pictorial representations with the user population. Many times people are confused by the given pictures and make wrong interpretations. What is essential, but lacking, are rules for guiding the design and evaluation of the non-verbal representations.

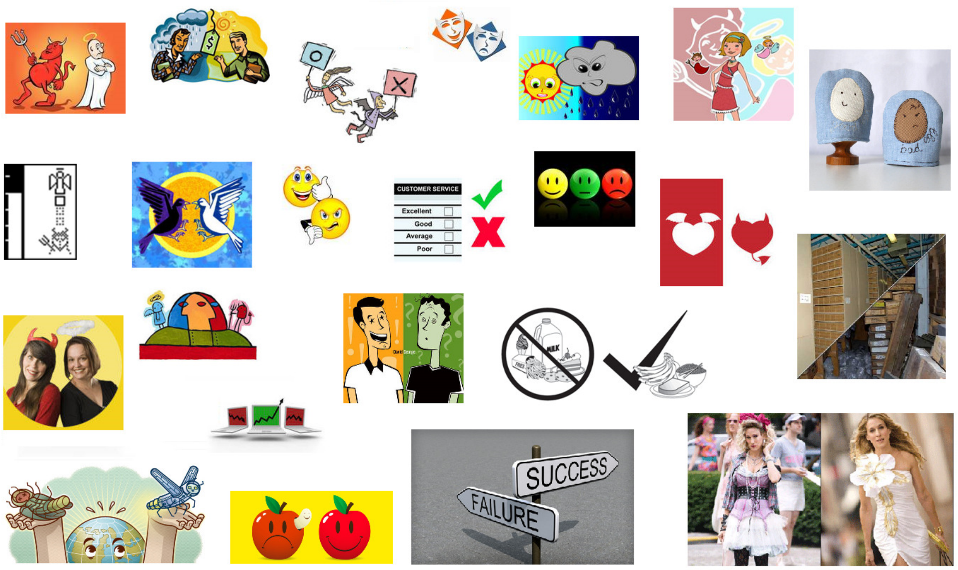


Figure 1-2. Example of the variance in representations (for the phrase "good or bad").

Third, many people find it hard to find a word given the current ways that the AAC vocabularies are structured. Categories, scenarios, and usage frequency are the most widely used vocabulary organizations, and they are not that intuitive in many cases (Figure 1-3). Most of the words are placed in a single location in the vocabulary. If the representation is not where he/she thinks it should be, the users may simply get confused. In one of the usability studies with Lingraphica, a participant became frustrated because he could not find the word "cat," because he never thought of "cat" being under the category of "people." Also, there are very few connections between words in different classes. People often need to traverse different parts of the vocabulary to locate words for a single sentence, and often they get lost in the process.

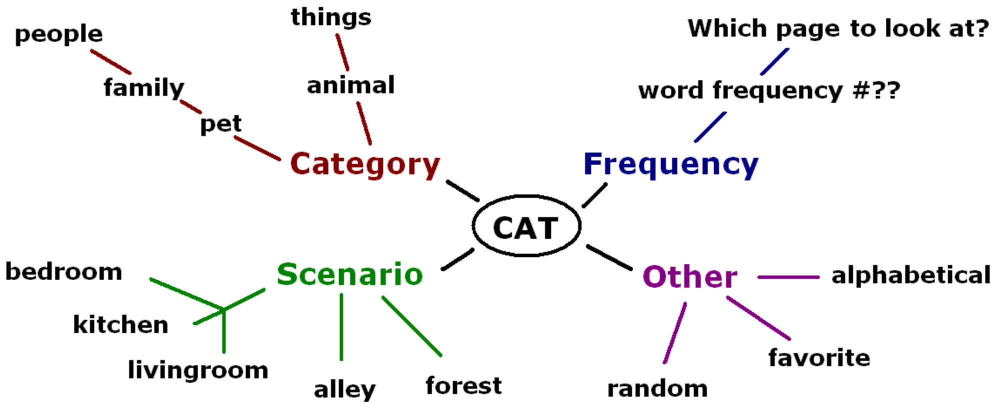


Figure 1-3. Example of vocabulary organization, where can I find the word "cat"?

Fourth, many of the AAC systems are implemented as a dedicated laptop-sized device, which people find too bulky to carry around. There is also stigma associated with using such devices for communication. Many people consider it embarrassing to pull out a big device, start up the software, and let it talks for him/her in the middle of a conversation. The latest AAC systems use mobile technologies, which are smaller, lighter and less intruding (and cooler). However, they run into other bottlenecks like limited storage and small screen size. Unlike a word dictionary, an AAC inventory requires a lot of space to store the media data, and a large enough display so that people can view the representations clearly. Small screen size also decreases the number of representations that can show up on one page, which adds more scrolling, clicking and context switching to the vocabulary navigation.

To address the problems in current Augmentative and Alternative Communication technologies listed above, research was conducted in 1) changing the non-verbal representation association from "with words" to "with concepts;" 2) exploring new media types as potential concept carriers, including web images, videos, and environmental sounds; 3) proposing methodologies for evaluation and refinement of multimedia representations; 4) investigating the Internet as a new platform for AAC systems. My research has shown that web images, despite their big variation in quality, are as effective as stylized icons in communicating all parts of speech to a healthy aging population as well as in communicating nouns to people with aphasia. Videos are the most effective stimulus for evoking verbs, especially with senior citizens. Environmental sounds, although not as good as visual representations in many cases, have shown significant advantage in illustrating certain types of concepts. The outcome of my thesis research is a novel AAC vocabulary in the form of a multimedia-enhanced semantic network (Figure 1-4), and a prototype online communication support system based on this effectiveness-verified vocabulary. A set of guidelines on the creation/selection of good multimedia representations is also proposed in this thesis. All of the work is different from prior research. The efficacy of this new multimedia communication support system on enhancing people's comprehension of information across language barriers has been demonstrated through studies with target populations in simulated information exchange scenarios.

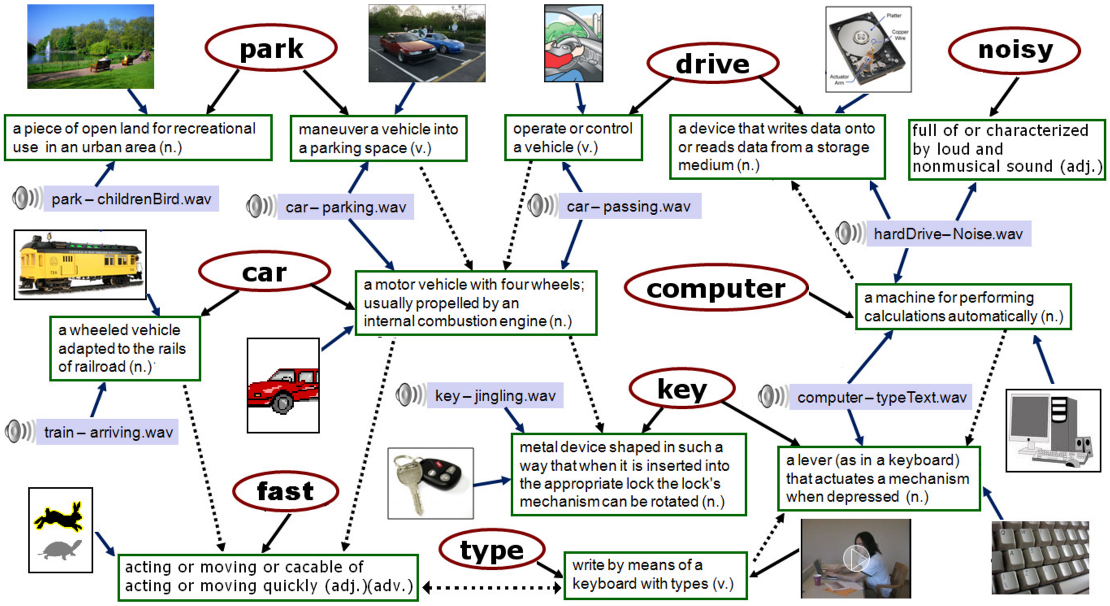

park
a piece of open land for recreational use in an urban area (n.)
park – childrenBird.wav
maneuver a vehicle into a parking space (v.)
car – parking.wav
drive
operate or control a vehicle (v.)
car – passing.wav
a device that writes data onto or reads data from a storage medium (n.)
hardDrive– Noise.wav
noisy
full of or characterized by loud and nonmusical sound (adj.)
car
a wheeled vehicle adapted to the rails of railroad (n.)
train – arriving.wav
a motor vehicle with four wheels; usually propelled by an internal combustion engine (n.)
computer
a machine for performing calculations automatically (n.)
key – jingling.wav
key
computer – typeText.wav
metal device shaped in such a way that when it is inserted into the appropriate lock the lock's mechanism can be rotated (n.)
a lever (as in a keyboard) that actuates a mechanism when depressed (n.)
fast
acting or moving or cacable of acting or moving quickly (adj.)(adv.)
type
write by means of a keyboard with types (v.)

Figure 1-4. Example of the multimedia-enhanced semantic network.

The rest of my thesis is organized as follows:

Chapter 2 Background Work reviews the literature in the usage of natural languages and the use of Augmentative and Alternative Communication techniques in the absence of linguistic communication. *Section 2.1 Language and Communication* summarizes the theories in how natural languages encode information, and how concepts and words are stored in and retrieved from the human brain. These theories provide an insight into how non-verbal representations, in place of words, should be constructed, organized and displayed for efficient use in communication. *Section 2.2 Language Disabilities and Language Barriers* points out the difficulties encountered in everyday life by people with impaired or limited language skills. It defines the target population of my thesis research with the important impact of the work highlighted. *Section 2.3 Augmentative and Alternative Communication (AAC)* provides detailed discussion of all the techniques in vocabulary generation, symbol system creation, and system design that have been employed by the existing AAC devices. Pros and cons of each technique are presented, and the remaining issues which hinder the practical usage of AAC technologies are brought up. This chapter defines the scope of my research.

Chapter 3 Research Overview: Online Communication Support via Multimedia summarizes the research problems revealed in the literature review chapter. Approaches to address these problems proposed in my thesis research are listed, followed by a summary of my research contributions and broader impact.

Chapter 4 Static Visual Representations for All Parts of Speech examines the evocativeness of web images in comparison to icons that are used in current AAC devices. *Section 4.1 Static Visual Representations: Perception and Usage* goes over the history of pictorial communication, introduces visual perception of static pictures, and proposes the validity of adding web images to AAC vocabularies. *Section 4.2 Automatic AAC Vocabulary Construction from Image Label Dataset* introduces a new unsupervised algorithm to automatically create image-concept associations for an AAC vocabulary from an image label dataset. The good performance of this algorithm proves the ease of construction and scalability of a web image-enhanced semantic network. *Section 4.3 Images vs. Icons: Evocativeness of Static Visual Representations* presents two studies that systematically compared the efficacy of web images and stylized icons as representations for concepts in different parts of speech. It is shown that images achieve similar performance as deliberately designed icons, and thus are valid to be included in an AAC vocabulary.

Chapter 5 Dynamic Visual Representations for Verbs moves on to explore the use of videos in conveying verbs in comparison against other stimuli. *Section 5.1 Difficulty in Visualizing Verbs and Proposed Approaches* discusses the problems with using static pictures to illustrate verbs which are often related to existence, actions, and change of status unfolded over time. Dynamic pictures, i.e. videos and animations, therefore are proposed as an alternative. The literature review on perception of dynamic pictures provides some theoretical grounds for this approach. *Section 5.2 Creation and Evaluation of Dynamic Visual Verbs* discusses in detail the generation of video representations for commonly used verbs. Two studies examined the effectiveness of videos as a visual verb stimulus against three other modalities (a single static image, a panel of four static images, and an animation). Videos were shown to be superior to the other stimuli in presentation precision, and thus are viable as a part of AAC vocabulary.

Chapter 6 Non-speech Auditory Representations for All Parts of Speech introduces a novel representation modality for concepts that has been under-investigated, i.e. environmental sounds. *Section 6.1 Auditory Perception and Sonification* presents the theories behind how humans perceive information related to sound-producing events from everyday sounds. It is the basis of the proposed approach on delivering frequently encountered concepts and messages via non-speech audio. Applications that use sounds to encode information are discussed. *Section 6.2 SoundNet Overview* introduces the overall idea of SoundNet, the environmental sound-enhanced semantic network as part of the multimedia vocabulary developed in my thesis research. Detailed information on the construction of SoundNet is provided. *Section 6.3 Evaluation of Non-speech Auditory Representations, a Labeling Study on Soundnail Precision* describes a large-scale sound tagging study conducted over the Internet that aims to evaluate the representation precision of non-speech audio clips. *Section 6.4 Evaluation of Non-speech Auditory Representation, a Soundnail Comprehension in Context Study* presents a second study that compared the information delivery efficacy of environmental sounds with that of AAC icons and animations. Although pictures in most of the cases achieved better comprehension accuracy, sounds outperformed the visual stimulus in about 10% of the concepts, especially those related to sound events and actions. These results suggest that non-speech audio is a viable and important alternative information communication channel.

Chapter 7 Multimedia Language Applications discusses applications aspects of the thesis research. *Section 7.1 Online Multimedia Language Assistant (OMLA)* introduces a web-based communication support system (OMLA) developed on top of the multimedia enhanced semantic network that is built through the research described in the previous chapters. OMLA provides an online popup multimedia dictionary with adaptable features that allows users to instantly view representations of unfamiliar concepts encountered during web browsing. The advantages of OMLA over text-based dictionaries, translations, and other online pictorial dictionaries are discussed. Two studies in food and medical care scenarios that investigated the performance of OMLA and multimedia representations on improving web information comprehension were conducted with non-native English speakers in comparison with native speakers. Study design and results are presented in *Section 7.2 Information Comprehension with OMLA in Food Domain* and *Section 7.3 Information Comprehension with OMLA in Medical Care Domain* respectively.

All the research methodologies and findings are summarized and future research directions are proposed in the last chapter, *Chapter 8 Conclusions and Future Work*.

CHAPTER 2 BACKGROUND WORK

2.1 Language and Communication

Communication is “the activity of communicating; the activity of conveying information (WordNet2010).” A **language**, by the definition of WordNet (WordNet2010), is “a systematic means of communicating by the use of sounds or conventional symbols.” A language is a system that encodes and decodes information with representations and a set of rules. There are many kinds of languages, from body languages to programming languages, from artificial languages (e.g. international auxiliary languages) to natural languages. This thesis focuses on natural languages and their usage in communication.

2.1.1 Natural Languages and Related Concepts

A **natural language**, e.g. English, Arabic and Chinese, is any language that evolves naturally from human mind. Articulate speech and the use of natural languages are one of the characteristics that distinguish humans from animals. Natural languages are used for communication, and they usually consist of a spoken component (speech) and/or a written component (writing). In a spoken language, the messages are encoded in speech sounds, while ideas are expressed through a symbol system in a written language. In the rest of my thesis, if not specified, the term *language* refers to a *natural language*. Also, if not specified, the examples are in one particular natural language, *English*.

Here are some basic terminologies and concepts regarding a language used in this thesis.

Concept, “an abstract or general idea inferred or derived from specific instances,” is people’s perception of entities in the world. **Meaning**, “the message that is intended or expressed or signified,” is the concepts carried in any communication. **Morpheme** is the “minimal meaningful language unit; it cannot be divided into smaller meaningful units.” A **word**, “a unit of language that native speakers can identify”, is the smallest free form that occurs in a language. A word can be a single morpheme (called free or unbound morpheme), such as “interest.” A word can also consist of several bound morphemes. For example, the word “uninterested” has three morphemes: “un-,” “interest,” and “-ed.” **Sense** is “the meaning of a word or expression.” Bigger language units can be constructed from words, such as a **phrase** (an expression consisting of one or more words forming a grammatical constituent of a sentence) and a **sentence** (a string of words satisfying the grammatical rules of a language). For example, “stay awake” is a phrase, and “I stayed awake till after midnight.” is a sentence. **Semantics** is “the meaning of a word, phrase, sentence, or text.”

Any terminology with "**semantic**" in it is discussing matters at the concept level. Another important aspect of a language, grammatical rules (or grammar) structure words into phrases and sentences. In linguistics (the scientific study of language), grammar refers to the set of logical and structural rules guiding the composition of language units. Issues associated with grammar are not involved my graduate research. The discussions in this thesis concern concepts, words (not bound morphemes), and on some occasions, phrases. Sentences only appear in the context of word sense disambiguation and comprehension evaluation.

The **lexicon** (or **vocabulary**), "a language user's knowledge of words," is the bridge between a language and the meanings expressed in that language. A lexicon is a collection of representations (words in the context of natural languages) mapped to concepts. A **lexeme** is a minimal unit (as a word or stem) in the lexicon of a language. For example, "run," "runs," "ran," and "running" are all members of the English lexeme "run." These different appearances of a word are called different **forms**. Among the members of the lexeme "run," "runs," "ran," and "running" are the **inflected form**, whereas "run" is the **base form**, or **stem** (or **root** or **lemma**) – the form of a word after all affixes are removed. The process of morphing a word from an inflected form to its base form is called **stemming**. One thing to note is that the inflected form of a word could be the base form of another, e.g. "glass" and "glasses (as the optical instrument)." To be more explicit, items in a lexicon are unique forms (strings of letters/speech sounds) – meaning (concepts) mappings. For example, the word "tree" has seven senses, which means there are seven units instead of one in the lexicon/vocabulary. These units are often referred as words but the definition of words in this sense is different from what was given in the previous paragraph.

2.1.2 Associations between Words and Concepts

A natural language uses units like words to express concepts. However, the relationship between words and concepts is not a one-to-one mapping (Figure 2-1). First of all, a word can have multiple meanings. For instance, "walk" can mean the act of walking or the path for walking. This phenomenon is called **polysemy**: the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings (according to WordNet). A word that has multiple senses is referred to as a **polysemous** word. The relation between senses of a word could be close (e.g. rain as the raindrops and rain as the event of water falling from clouds) or far apart (e.g. park as a recreation area and park as the act of placing a vehicle in a spot). The distance between meanings is called **semantic distance**. Usually, the sense of a word can be identified by the **context** (discourse that surrounds a language unit and helps to determine its interpretation) around it, such as the other words in the same sentence or surrounding sentences. For example, a "bat" could be a club used for hitting a ball, and could be the nocturnal mouselike mammal. The word “bat” most likely takes the former sense in the sentence “I swung the bat but missed the ball.”

On the other hand, a concept can be encoded in various word forms. For instance (Figure 2-1), the words “walk,” “walkway,” and “paseo” can all be used to refer to the concept of “a path set aside for walking.” Other examples include “child” and “kid” as a young person of either sex; “put” and “place” as (verb) put into a certain place or abstract location; “slender” and “slim” as (adjective) being of delicate or slender build; and “quickly” and “rapidly” as doing something (adverb) with speed. Words that denote the same meaning can be interchanged in many contexts involving this meaning. They are called **synonyms**. In WordNet, the group of words for the same concept forms a unit called **synset** (synonym set). There are also occasions when words with similar but not exactly the same meanings are used in place of one another in certain contexts, for example, a “tasty (pleasing to the sense of taste)” meal and a “delicious (extremely pleasing to the sense of taste)” meal. The slight differences could be the aspect that is being emphasized, or the degree/extent, or formal vs. less formal. However, there are words having literately very similar or even the same meanings but should not be used as substitutes for one another, due to subtleties such as positive or negative emotional coloration. For example, “provoke” has a more negative color than “evoke.”

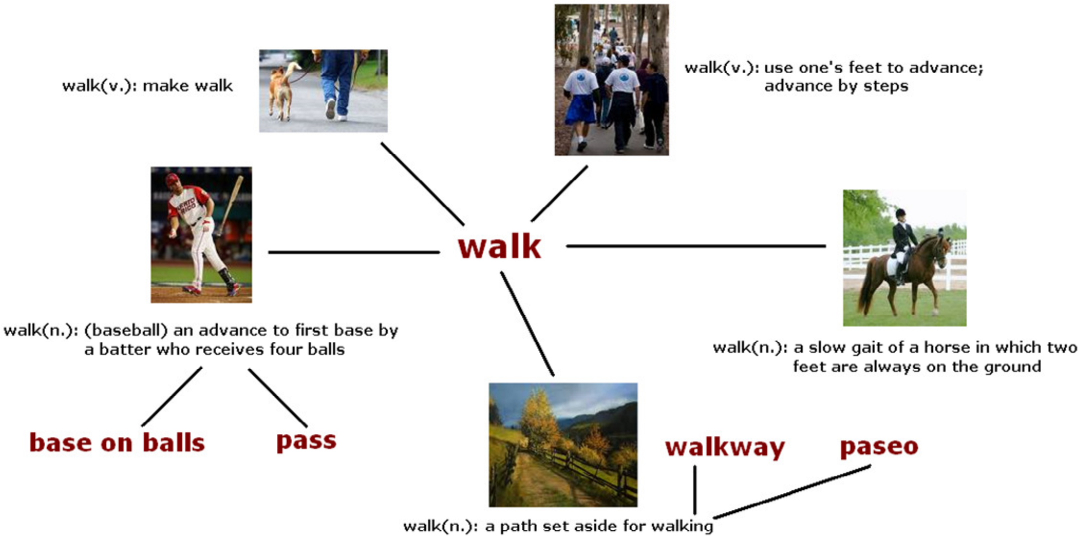

walk(v.): make walk
walk(v.): use one's feet to advance; advance by steps
walk
walk(n.): (baseball) an advance to first base by a batter who receives four balls
walk(n.): a slow gait of a horse in which two feet are always on the ground
base on balls
pass
walkway
paseo
walk(n.): a path set aside for walking

Figure 2-1. Example of multiple-to-multiple mapping between words and concepts (“walk”).

2.1.3 Mental Lexicon: Language Storage and Access

Before discussing how a language is used for communication, I briefly discuss how the human mind stores and retrieves language units such as words. The **mental lexicon** (Oldfield1966) refers to a person’s internalized knowledge of the properties of words, i.e. how humans store and manage words in the mind, and access words they intend to express. Mental Lexicon is different from a physical dictionary. A dictionary usually lists items in linear indexing (e.g. by alphabetical order), and the search speed for finding a word is independent of its linguistic properties. For example, it is less likely that people will find "I," one of the most commonly used words, and "iamb." a word much further down the frequency list, on the same page in a dictionary. On the contrary, familiarity effect is detected in human lexicon access. If asked to name a word starting with a given letter, a person almost always comes up with a high-frequency/high-familiarity word. For example, people are much more likely to say "ice" than "iamb" when asked for a word that begins with the letter "i," even if it is against the alphabetical order. It suggests that the structure of the mental lexicon is different from a physical dictionary. A speaker can recognize a word of his/her native language in 200ms or less, and reject a non-word sound sequence (e.g. a made-up word or a mispronunciation) in about 0.5 second. Given the size of a person's working vocabulary (about 20,000 word form-meaning pairs), this speed is extraordinary. Mental lexicon must be organized in a systematic way to allow such efficient word finding.

2.1.3.1 Free Word Association Norms

Psychologists have been studying the magic of mental lexicon for hundreds of years. The way they have been approaching the problem is by collecting **free word association norms** (Moss and Older1996). To gather information about what words people consider closely related, psychologists ask people to respond to a stimulus word with the first word they think of, as quickly as possible. Although both the stimulus words and the response words are polysemous, free word association norms give some insight as to what words are commonly used together. Experiments and surveys on free word association norms have been conducted with people of different age, gender, race, culture, nationalities, and language background, and even with mentally ill (not severely) individuals. The results are fairly consistent across all subject groups. It suggests that people have quite uniform way of associating words and organizing their mental lexicon.

By analyzing the word pairs in the norms, it is discovered that most of the stimulus words and response words are connected through meanings. For example, given the word "toe," a third of the people come up with "foot" as their first response. Some of the other words that are generated due to semantic closeness are "nail (23%)," "shoe (5%)," and "big (2%)." There are two other associations that people use to connect words together, i.e. by pronunciation/speech sound similarity and by context/experience. For instance, if the stimulus word is "explicit," some people come up with "implicit" because the two share some syllables. Examples for relating words based on context or personal experience are "toe" and "ballerina," "toe" and "bite" (the person's toe got bitten by a dog once).

Free word association norms suggest that words are stored in clusters in mental lexicon, mostly by connections in their meanings. It is similar to how people arrange books in a bookstore. To facilitate efficient search, books are usually grouped into shelves by areas (e.g. fiction, non-fiction, bibliography, science, art, etc.), by topics (e.g. mystery, romance, etc.), by authors (e.g. within mystery one may find Arthur Conan Doyle, Agatha Christie, etc.), by year of publication, or by alphabetical order of the book title. We seldom see books that are put next to one another because of the similar color of the cover, the size, or the number of pages. It is even less possible to scatter books around randomly. Likewise, the underlying mental lexicon organization that keeps related terms in adjacency allows people to generate speech and text from tens of thousands of options in response to the topic and context of communication within less than a blink. Many models that try to explain the way mental lexicon stores and retrieve words and meanings have been proposed.

2.1.3.2 Semantic Memory and the Activation and Spreading Theory

When words are kept in a human mind, what is stored is more than just the word forms. It includes semantic (meaning), phonological and orthographic (speech sound), as well as morphological (word form) information for each word. Various models have been developed to explain the word recognition process. The Wernicke-Geschwind model for example, separates the process of speech sound and meanings. It suggests that the Wernicke's area in the human brain transcribes auditory input to phonological representations from which the semantic network in the mental lexicon is accessed. Some of the models look into how word forms or phonological/orthographic information are mapped to concepts. The Cohort Model (Marslen-Wilson1978) is a parallel activation model. In this model, a large number of hypothesized words are triggered by a speech input. As more auditory information rolls in, activation of the words that are concluded to be irrelevant ceases. The Interactive Activation Model (McClelland and Rumelhart1981) uses letters and their positions in the word as the stimulus for word activation and selection. And later the TRACE model (a kind of connectionist model) was formed adding another layer for phoneme processing (McClelland and Elman 1986). In the TRACE model, broader lexical selection is influenced by sentence context and then the sub-lexical selection is affected by word-level context. The Neighborhood Activation Model (Luce and Pisoni 1998) proposed that words are grouped in to neighborhood based on their phonetic similarity in sound patterns. Once a speech input comes in, the neighborhood with the most similar auditory pattern gets activated. The words in the neighborhood are further discriminated by the degree of activation according to their phonetic similarity. A similar idea has been applied to modeling process and search in the semantic memory.

Semantic memory is the part of human long term memory that stores context-independent concept-based knowledge. That is to say, this knowledge, extracted from general understanding of the world, is not bound to any particular person or place or experience, for example, “a whale is an animal not a metal.” Concepts contained in the semantic memory are interlinked by the knowledge. As in the example above, the concept “whale” is associated with “animal” but not “metal.” The associations are directed with different strength. For example, when people are asked to name an animal, a whale is less likely to be the top one, whereas “dog,” “bird,” and “fish” have higher possibility to be mentioned. This phenomenon is summarized in the prototype theory (Rosch1978). Prototype theory says that certain members of a category are closer to the center than others, which suggests that the distances between connected concepts are not equal. There are categories, and one or more members are representative of the category.

Several models have been created to simulate the semantic memory, such as the semantic feature-comparison model (Smith et al.1974), the search of associative memory model (Raaijmakers1981), and the statistics model like the latent semantic analysis model (Landauer1997). One of the most popular models for semantic memory is the semantic network model (Collins and Quillian1969). In the semantic network model, concepts are represented as interconnected nodes (Figure 2-2). The distance between nodes indicates the concept similarity. The sense of a word can be determined in terms of the other words linked to it. For example, the word “blue” in Figure 2-2 means the color intermediate between green and violet, not the gloomy emotion.

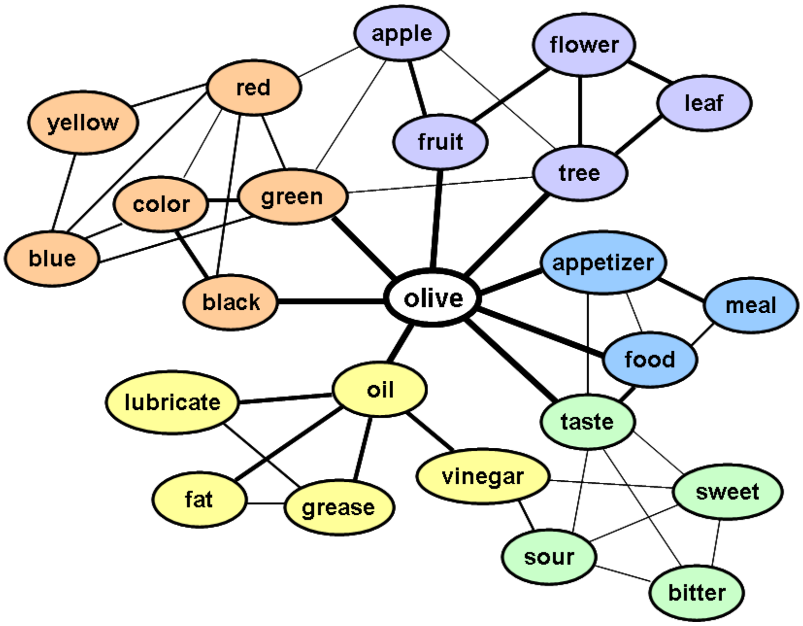


Figure 2-2. Example of activation and spreading in semantic network.

The spreading-activation model (Collins and Loftus1975) describing how concepts are process is based on the semantic network model. The basic idea is that once a concept becomes active, the activation will spread via links to other nodes. The effect of the activation diminishes as it travels further down the network. In Figure 2-2, the thickness of the lines between nodes represents the degree of activation spread from “olive.” The access speed for the current node to the other activated nodes is faster than the inactive ones. For example, the reaction time for people to say “oil” where hearing “olive” is shorter than say “blue.” Free word association norm reflects the semantic network and spreading-activation models to some extend. WordNet, which is mentioned earlier, is built on top of these models.

Once a concept gets activated in semantic memory, the word form of the concept must be assembled in order to be expressed in speech or writing. If the second step is incomplete for some reason, the person is experiencing what is known as the tip-of-the-tongue (TOT) phenomenon. TOT is a situation when people know the under-lying concept but fail to retrieve the word for it. A common scenario is when people cannot recall the name of a known person. People with language disabilities or language barriers can be viewed as having persistent tip-of-the-tongue experience (although not the same mechanism is in place). For example, when a foreign tourist wants to find something in a store, he knows what he is looking for, but does not know the name for it in the local language. In this case, there is no existing link between concepts and the word form becomes inaccessible, because the link, itself, does not exist.

2.1.4 Semantic Relationships

Knowing that related words are stored nearby in the mental lexicon, researchers would like to find out more about what exactly these relationships are. Synonym is only one of many semantic relationships between concepts (word senses). Semantic relationships (also called semantic relations in some documents), describe how concepts are connected to one another. Figure 2-3 shows the semantic relationship taxonomy (based on Storey1993). **Antonym** means a word expressing an opposite meaning to the given word (e.g. good and bad). **Hypernym** means a word that is more generic than the given word. **Hyponym** means a word that is more specific than the given word. **Meronym** means a word that is part of a larger whole. **Holonym** means a word that is the whole of which the given word is a part. **Instance** means a word that is a special case of the given word. In WordNet, an instance, referring to named entities, is a specific kind of hyponym. **Troponym** (for verbs) means a word is in a special manner of the given word (e.g. yell and talk), which is a kind of hyponym. **Entailment** (for verbs) means one event necessarily entails another (e.g. snore-sleep) If two concepts share the same hypernym, they are **sister terms**. The sister term of the hypernym of a given word is called its uncle term, and the hyponym of the sister term of a given word is called its nephew term. Examples are shown in Figure 2-4.

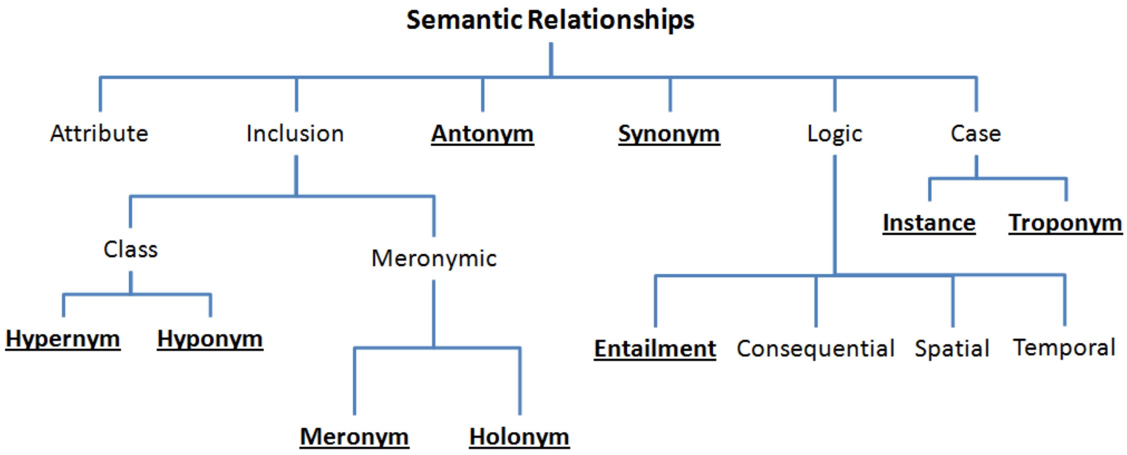


Figure 2-3. Taxonomy of semantic relationships, the underlined relations are used in WordNet.

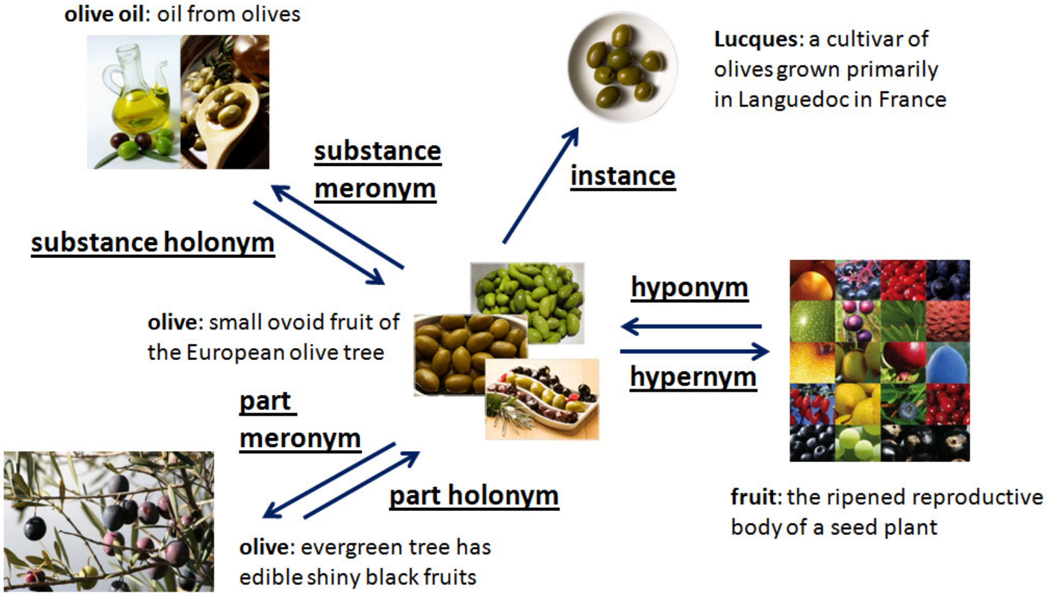

olive oil: oil from olives
Lucques: a cultivar of olives grown primarily in Languedoc in France
substance meronym
instance
substance holonym
hyponym
olive: small ovoid fruit of the European olive tree
hypernym
part meronym
part holonym
fruit: the ripened reproductive body of a seed plant
olive: evergreen tree has edible shiny black fruits

Figure 2-4. Examples of various semantic relationships.

WordNet (which has been mentioned multiple times) (Fellbaum1998), is an online lexical network developed at Princeton University. It provides separate semantic relationships for nouns, verbs, adjectives and adverb. These relationships are highlighted in Figure 2-3. It is a great resource for determining how words and concepts are linked to one another. However, the relationships across different parts of speech (e.g. between nouns and verbs) are largely missing, and there are other important connections between concepts that are not included in WordNet. Therefore, there have been other attempts to explore other semantic relationships. **Evocation** (Boyd-Graber et al.2006) (Nikolova et al.2009), for example, is a semantic measure that rates how much one word brings to mind another. Evocation gives a directed weighted score, which could be across different parts of speech. Frame semantics (Fillmore1982) connects linguistic semantics to empirical knowledge. Frame semantics groups concepts, which are inevitable in order to understand any one of them, into a **Frame**. A frame denotes a cognitive scheme, an experience or knowledge. For example, "seller," "sell," "buyer," "buy," "goods," and "money" occur in the same frame "Commercial Transaction." **FrameNet**, which is an online lexical resource built in Berkeley, is based on the idea of frame semantics. FrameNet relates concepts based on encyclopedic knowledge. Research has also studied word associations through **occurrence** in corpus data.

A semantic network of concepts can be constructed based on the relationship mentioned above. The distance between nodes (i.e. concepts) can be determined. Semantic distance is also referred to as semantic relatedness or semantic similarity. The closer two concepts are in the semantic space, the more similar their meanings are.

2.1.5 Linguistic Properties

Words, together with their meanings, have various linguistic properties. **Part of speech** (also referred to as word class), "one of the traditional categories of words intended to reflect their functions in a grammatical context," is one of the most important properties. Traditionally, words are classified into eight parts of speech: **noun** (a person, place, thing, quality, or action), **verb** (an action, occurrence, or state of existence), **adjective** (an attribute of something), **adverb** (a word that modifies something other than a noun), pronoun (a function word that is used in place of a noun or noun phrase), preposition (a function word that combines with a noun or pronoun or noun phrase to form a prepositional phrase), conjunction (an uninflected function word that serves to conjoin words or phrases or clauses or sentences), and interjection (expressing emotion and having no grammatical relation with other words in the sentence). Nouns, verbs, adjectives, and most adverbs are words to which an independent meaning can be assigned. They are called **content words**, while the rest of parts of speech are referred as **function words** (or grammatical words), which are uninflected and serve a grammatical function but have little identifiable meaning. For instance, in the sentence "Ouch, a mosquito just sneakily bit me on my right arm," the content words include "mosquito (n.)," "arm (n.)," "bit – bite (v.)," "right (adj.)," and "sneakily (adv.)," while the rest are function words. My thesis research only focuses on content words.

Different parts of speech are acquired at different times during language learning. A lot of research (Gentner1982) has shown that nouns have significant advantage over verbs in the **age of acquisition**, despite the linguistic characteristics of the language (e.g. verb-second as English or verb-final as Korean). Nouns and verbs both appear earlier in the vocabulary of a child than other content words and function words. In general, words that are learned earlier are recognized more quickly and used more often. Loss of vocabulary follows a similar pattern as age of acquisition (Bleser2003). It has been shown that people with language disorders (e.g. aphasia) may have lost all pronouns, articles, and conjunctions, although they may retain a certain amount of nouns and verbs (Carroll 1999). Also, the impairment of verbs is often more severe than that of nouns for this group of people.

Word frequency is a property indicating how often a word appears in general language usage. It reflects to a certain degree, the importance of words in language usage. Research has shown that highly frequent words are likely to be acquired earlier and retained longer than words that are less common (Bird2001). **Familiarity** is a measure of personal knowledge or information about a given word. Although in general people are more familiar with the most frequently used words, the familiarity rating may change from person to person due to domain knowledge, life experience, and topics of interests.

Concreteness is defined as the ability to see, to hear, and to touch something (Paivios1968) (Richardson1975). A concrete word refers to objects and events that can be seen and/or touched. An abstract word denotes a concept that has no physical embodiment. More general terms are not necessarily more abstract (e.g. animal). **Imageability** (or imagery) refers to the ability of a word to evoke a sensory mental image (Richardson1975). Imageability is highly correlated to concreteness. If a concept has some physical reference, it is likely that people can picture it in their head (Paivio1968).

All of these properties, especially parts of speech, concreteness, and imageability, may have an effect on how concepts are linked (Altarriba1999) and processed by the human brain (Swaab2002), and thus, are studied in my thesis research.

2.1.6 Language Usage

Having the ability to retrieve words for intended ideas, people use natural languages as the principal means to communicate. Language usage in everyday spoken communication can be divided into three categories: small talk, topical conversation, and storytelling. Small talk (chitchat) is a light informal conversation for social occasions, such as a greeting in the hallway. There is no actual information exchanged or any functional topics involved. Conducting small talk is more for a social purpose, for example, at the beginning or the end of a conversation, or as a silence-breaker. Topical conversations are the exchange of information and/or ideas around certain topic. A topical conversation involves short speech (usually a couple of sentences) going back and forth between multiple conversation partners. Storytelling refers to a long narration of people, events, or experiences.

Language usage patterns vary by age and gender. (Stuart1991) has shown that elderly people tend to spend less time engaging in small talk. The percentage of storytelling in their daily language usage increases, since older adults tell stories to maintain social bonds with peers, to educate the young, and to pass down knowledge and traditions. The topics of interest have changed as people aged as well. For example, on average people spend half of their time talking about the present, while for people over the age of 75, about 50% of the topics refer to the past. In addition, females and males use language differently. Distinction in conversation topics is one of the examples. Male topics include work, money, taxes, legal issues, army experiences, sports and amusements; while fashion, shopping, household needs, etc. are found more in conversations among women. Women may also be taught to speak in a different manner (Lakoff1977).

En sum, this section introduced the background information on natural languages, the association between words and meanings, various relationships between concepts, storage and access of concepts and words in the human mind, and language usage in communication. The theories and models about semantics and vocabulary organization discussed above serve as the theoretical basis for my thesis research.

2.2 Language Disabilities and Language Barriers

The target population of this thesis research is introduced in this section, including people with language disorders (disabilities) and people with language barriers.

2.2.1 Communication Disorders and Language Disorders

Communication Disorders (including speech disorders and language disorders) are problems in communication that results in difficulty in producing correct speech or inability to understand or use languages. Examples of causes of communication disorders include hearing disorders and deafness, voice and speech problems, developmental disabilities, learning disorders, autism, and language disabilities (MedlinePlus, 2010). If a person has trouble with his/her voice or with producing speech sounds correctly or fluently, he/she has a speech disorder. If a person has trouble understanding others (speech or written materials) or sharing his/her thoughts and feelings, he/she has a language disorder. In my thesis research, I am focusing on one kind of communication disorders, i.e. **language disorders (or language impairments)**.

Language disorders (Lahey, 1988) involve difficulty in processing natural languages. The disorders can affect the processing of linguistic information such as semantics (the meaning of the words), morphology (the structure of words), and/or syntax (sentence structure). It can result in disabilities in any combination of language comprehension (**receptive language disorders**) and language production (**expressive language disorders**). Spoken language or written language or both can be impaired. Language disorders can appear in sign languages as well. Expressive language disorders (impaired vocabularies and grammars) are more common among children. The cause of developmental expressive language disorders remains unknown, while acquired expressive language disorders result from brain damages. Receptive language disorders (inability to understand languages) are sometimes referred as learning disabilities. Examples of language disorders include aphasia, echolalia, mixed receptive-expressive language disorder, and specific language impairment. My thesis research focuses on **aphasia**, the most common language impairment.

2.2.2 Aphasia and Its Social Impact

Aphasia is “an acquired communication disorder that impairs a person's ability to process language, but does not affect intelligence.” (National Aphasia Association 2010). Aphasia is most commonly acquired after a stroke (25~40% stroke survivors). Other causes include brain tumors, head injuries (30% of the severely head injured people), and neurological diseases. Currently, about one million Americans are affected by aphasia, and the number is increased by more than

100,000 each year. Aphasia may occur in people of any age, gender, race and nationality; however, it is most common among elderly people and the risk increases by 1~7% for each advancing year (15% of individuals under the age of 65 had aphasia and the percentage is 43% for people 85 and older) due to the bigger possibility of having a stroke (Engelter et al.2006).

Aphasia can lead to deficiency in any language modalities (Goodglass1993). The variation in aphasia symptoms comes from the differences in the damaged regions in the language center (almost always located in the left hemisphere of the human brain). Some of the varieties are global aphasia (the most severe aphasia, where patients can neither read, write, speak, nor understand speech), Broca's aphasia (patients are non-fluent in speech and writing but fine with comprehension), Wernicke's aphasia (patients are limited in reading, writing, and speech understanding, but can still produce connected speech), and Anomic aphasia (patients show no difficulty in comprehension, but are unable to retrieve the right words for intended meanings). No existing drugs or medication can cure aphasia, and surgery can only target certain occasions with brain tumors. If symptoms of aphasia last over two to three months after acquirement, a full recovery is unlikely. Through speech and language therapy, some people’s condition can improve over the years, but it is a long process.

People with aphasia encounter great difficulties and frustrations in communication, and they are often withdrawn from social activities. According to a survey conducted by the National Aphasia Association, 72% of the aphasic respondents lost their ability to work. Those who reengaged in certain occupations reported big constraints due to their language disabilities. Aphasic people’s social connections have been severely affected as well. Although their intelligence remains intact, 70% of the individuals with aphasia participated in the survey felt that people, even family members, avoid contact with them either because of the difficulty in communication, or them being mistaken as mentally retarded. It is hard to maintain friendship with the lack of timely communication on common interests as well as the lack of humor in speech (Davidson et al.2008). Some research (Kauhanen et al.2000) (Code et al.1999) showed that 60~70% of the people developed depression 12 months after they acquired aphasia, and 33% of these people got worse. Aphasia has negative impacts on many aspects of the quality of people’s life.

2.2.3 Language Barriers

Not only individuals with language disorders experience difficulty in communicating through natural languages, the existence of language barriers makes verbal communication ineffective as well. A language barrier is generally referred to the circumstance when people who do not share a common language try to communicate with one another. People facing language barriers have trouble understanding each other and expressing their own thoughts, which is similar to the situations encountered by people with aphasia. The ultimate way to resolve a language barrier is to learn each other's language. It takes a lot of time and effort, and is not realistic if people just travel to a foreign country for a short period of time, or in the context of short-term international collaboration. Furthermore, learning a new language is more challenging and cumbersome for people moving to a new language environment at an adult age. Language barriers have been shown to be a formidable obstacle to the social welfare of immigrants, including health care, education, and employment (Kretsedemas2005). Take the United States as an example, 47 million (15%) of the US residents speak a language other than English at home. The percentage of population with limited English skills is 13% in New York, 14% in Texas, and 20% in California. On the contrary, over 50% of the hospitals contacted in a study conducted by the Association of Community Organizations for Reform Now had no access to a single Spanish-speaking staff. Although not as severe as aphasia, language barriers hinder people's well-being.

2.2.4 Bridges to Communication Gaps

People have explored many ways to facilitate information exchange, and to make up for the loss of verbal communication. The simplest methods are those used naturally as one talks: informal vocalizations, eye gaze, facial expressions, hand gestures, and body postures (Beukelman and Mirenda2005). One can point to things or mimic. People have been seeking external aids as well (Scott1998). For example, "paper and pen" – sketching drawings and words (if available); flash cards and picture/translation books – showing pictures of intended meanings; and speech partners. As technologies advance, computers and electronic (high-tech) devices have been introduced as a communication assistant. Compared to the simple methods and low-tech (non-electronic) aids, electronic devices have bigger storage for vocabularies, more display and interaction modalities for better accessibility, ease of personalization, and are less likely to be emotional or tired (Sandt-Koenderman2004). However, high-tech systems may have mobility issues with a longer learning process and possible stigma in usage. The all of the methods and technologies mentioned above, especially those designed for people with impairments or restrictions on the production or comprehension of spoken or written languages, are called **Augmentative and Alternative Communication**.

2.3 Augmentative and Alternative Communication (AAC)

In the field of Assistive Technologies, Augmentative and Alternative Communication (AAC) refers to any attempt that "when necessary compensates for temporary or permanent impairments, activity limitations, and participation restrictions of individuals with severe disorders of speech-language production and/or comprehension, including spoken and written modes of communication (American Speech-Language-Hearing Association)." It can be generalized to any method of communication that supplements or replaces speech and handwriting (e.g. gesturing to friends across the table in a noisy bar, pointing to the menu when one cannot pronounce the name of the dish). The key idea of AAC is to fully utilize the remaining abilities of people with communication difficulties, in order to compensate and/or bypass the missing language connection (Millar and Scott 1998). Everyone benefits from AAC one time or another, while people with language disorders (e.g. aphasia) rely on AAC all the time as the main approach to reaching out to the world. AAC can be applied in all sorts of settings, from basic conversations with family and friends, work or school, to accessing professional services like medical care (Flippo et al.1995). AAC can promote people's access to welfare facilities and services, can enhance the independence of people with language disorders or barriers, can re-integrate them into the society (McFadden1995), and can improve people's quality of life (Macdonald1994).

Any AAC methods or techniques consist of three important components (Kangas and Lloyd1998): a system of representing meanings (words, speech, or pictorial symbols), a communication medium (e.g. body, paper, or computer), and the way to access the representations and transmit the message. An AAC system is an instance of the integration of these three components. My thesis research interests in computer-based AAC systems, and thus the discussion will concentrate on the representation system and means of access – the AAC vocabulary. Any AAC system needs to answer four essential questions: 1) what words and phrases should be included in a functional and motivational AAC vocabulary, i.e. **vocabulary selection**; 2) what stimuli should be used to convey meanings, i.e. **vocabulary representation**; 3) how the stimuli should be displayed for easy navigation and retrieval, i.e. **vocabulary organization**; and 4) how the stimuli should be accessed and what is produced, i.e. **vocabulary access**.

2.3.1 AAC Vocabulary Selection

To establish a collection of representations of messages, one first needs to identify what words are to be encode. The simplest strategy is to include words with the highest occurring frequency in common communication patterns in the AAC vocabulary. However, language usage and topics of interests vary from individual to individual, and thus, a generalized vocabulary may not satisfy personal needs of the AAC users. It also means that a vocabulary customized for one person may not be suitable for another. This paradox suggests that an AAC vocabulary should best reflect a common interest of the target user population, with the flexibility for extension and customization (Beukelman et al.1991). The (initial) choice of an AAC vocabulary should be sensitive to user profiles, communication environments and partners, cognition and language ability/disability, literacy, and the balance between the core vocabulary (85%~95% of what is said) and the fringe (personalized) vocabulary (5~15% of communication) (DynaVox2010). Some of the topics that have been identified to be important for daily communication are small talk (e.g.

greeting), family and friends, daily routines, food and medication, seeking help (e.g. go to the bathroom), transportation, transaction (Beukelman and Mirenda 1992).

Several methods for vocabulary selection have been studied: search in dictionaries and word lists based on frequency and usage pattern (Beukelman and Mirenda1998); category inventory, such as people, places, feelings, etc. (Carlson1981); Ecological/Environmental Inventories (Carlson1981); documentations of usage scenarios, such as diaries (Beukelman and Mirenda1998), storytelling (Fried-Oken 1995), and scripts of dialogs (Glennen and DeCoste, 1997); and interviews with AAC users and their support network, e.g. family, friends, and caregivers (Yorkston, 1989). How the vocabulary is collected has direct impact on the choice of display, which will be demonstrated in the vocabulary organization section.

2.3.2 AAC Vocabulary Representation

Once an inventory of words and phrases is established, a representation system (or **symbol system**) needs to be constructed to illustrate the meanings of each term in the vocabulary. In a natural language, the symbols are the letters and words, while in AAC, gestures/hand signs, line drawings/icons, images/photos, and tangible objects can all be employed as a symbol supplementing or replacing words (Mirenda2003). The first important thing to determine is how the symbols are mapped to a single meaning, i.e. the **coding mechanism**.

2.3.2.1 Coding Mechanism

Three approaches have been explored: alphabet-based methods, single meaning picture, and semantic compaction.

(1) The **alphabet-based** methods still use characters as the symbol unit, and it is suitable for people who retain literacy skills to assess the vocabulary through spelling. Four strategies fall into this category: whole word, full spelling, letter encoding (abbreviation-expansion), and word prediction (Figure 2-5). Using the whole word method, the entire words are displayed for selection. Using the full spelling method, only letters are given, and the users have to spell out a word by selecting the whole sequence of letters. Using the letter encoding method, the common words are encoded in an abbreviation with 2~3 letters, and in order to retrieve a word, people can either type in the abbreviation combination or select it on the screen. Using the word prediction method, the system will automatically guess what the word is as people start to spell it out. A suggestion list is given based on the prediction and the user can pick out the right one.

The alphabet-based methods have the advantage of a small symbol set (alphabet). In many languages, any word can be spelled with 20 to 30 letters plus a space character and several punctuation marks. The disadvantages of an alphabet-based method are the requirement of literacy and learning, and the number of selecting operations (the length of the symbol sequence), especially in full spelling. For example, the average English word length is five letters, and thus on average five symbols need to be clicked on to compose a word.

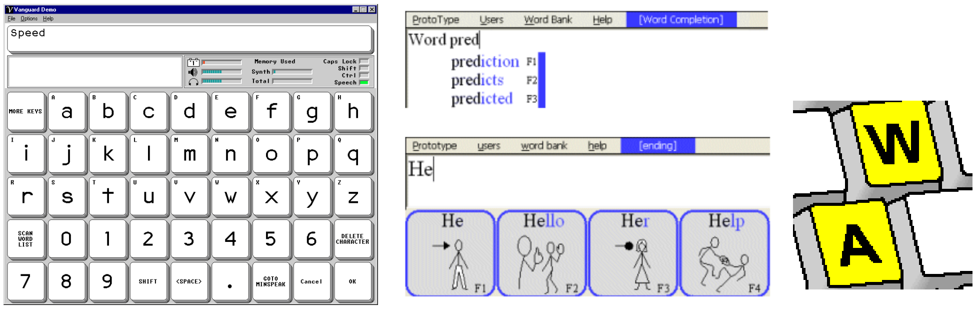


Figure 2-5. Alphabet-based methods: full spelling (left), word prediction (middle) and letter encoding for the last name “Allwright” (right).

(2) The **single meaning picture** method uses a picture to represent each individual word, and is normally the first choice to depict people, places, and objects (Figure 2-6). For each word, only one selection is required, and the length of symbol (picture) sequence equals to the length of the words in the sentence or phrase. Furthermore, the single meaning picture method does not require literacy skill. That is to say, people can still retrieve a word without knowing its actual spelling. However, the pictures must be able to evoke the assigned meaning easily, quickly, and accurately, and fine distinctions should be made for similar ideas, e.g. give, bring, take and accept. A picture may evoke several meanings (e.g. a picture of a car on the road may evoke “car,” “road,” “drive,” and “traffic”), and can introduce ambiguities. Another shortcoming of the single meaning picture method is the size of the whole symbol set, which is basically equivalent to the number of words and phrases in the vocabulary. Studies show that the working vocabulary contains 1,000 words for a normally developing 3-year old child, 2,500 to 5,000 words for a 5 to 6-year old, and about 20,000 words for a typical college graduate (Wren2003). It means 1000 pictures are needed to cover a 3-year old child’s vocabulary, and at least 2000 picture are required to illustrate the 2000 most common words in order to cover 96% of the materials appearing in the oral communication context (Schonell1956).

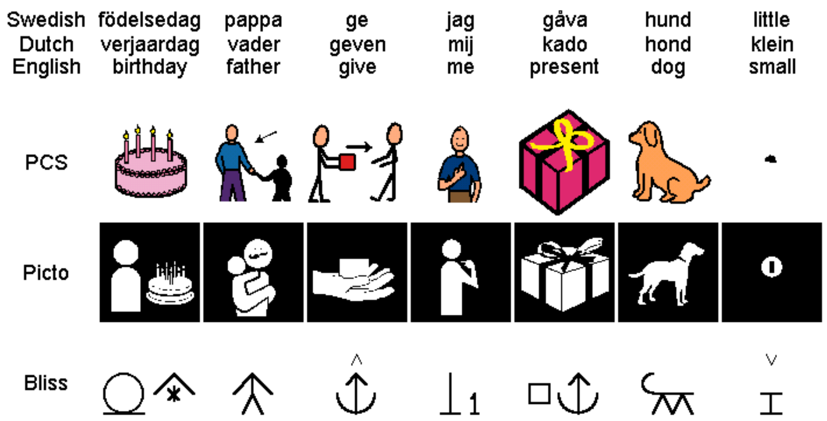

Swedish födelsedag pappa ge jag gåva hund little
Dutch verjaardag vader geven mij kado hond klein
English birthday father give me present dog small
PCS
Picto
Bliss
1

Figure 2-6. Examples of the single meaning picture coding method in comparison to natural

(3) The third coding method is called **semantic compaction**. In contrast to the single meaning picture method, semantic compaction uses multiple meaning pictures (also called icons) in short sequence to represent a single word or phrases (Baker1986). It is based on the observation that usually a picture can evoke more than one meaning, for example, when a person sees a picture of a glass of juice, he/she might think of the noun "juice," the verb "drink," and the adjective "sweet" or "thirsty." Semantic compaction applies the idea that a combination of pictures can help differentiate the intended meaning. The sequence consists of a key picture (where the meaning is derived from) and a set of functional pictures that defines the scope of the meaning, e.g. part of speech and category. The illustrated meaning changes when some of the pictures in the sequence or the order of the sequence changes (Figure 2-7). Minspeak (Minspeak2010) has been designed based on this method.

Semantic compaction does not require literacy (spelling of the words), but since it inherits the natural language structure (parts of speech and category information), users should maintain certain language skills. Also, training is necessary for the set of rules of picture sequencing, as well as the recognition/memorization of the functional pictures (e.g. the paint brush pictures represents *adjective*). Although semantic compaction utilizes the polysemy of a picture, it still has the ambiguity in the sequence. For example, an apple picture with a noun picture could be "apple," "fruit," and "food"; and with an adjective picture could be "red," "sweet," "juicy," "delicious," and "hungry." Compared to the single meaning picture method, semantic compaction has a much smaller symbol set, since the number of functional pictures is fixed, and only a small number of key pictures are necessary. It is the variety of combinations that gives it the power to cover a big vocabulary. The semantic compaction method also involves less selection operations than the alphabet-based methods. Studies (Hill2004) showed that semantic compaction had the best performance in accessing the core vocabulary among the three coding mechanisms.

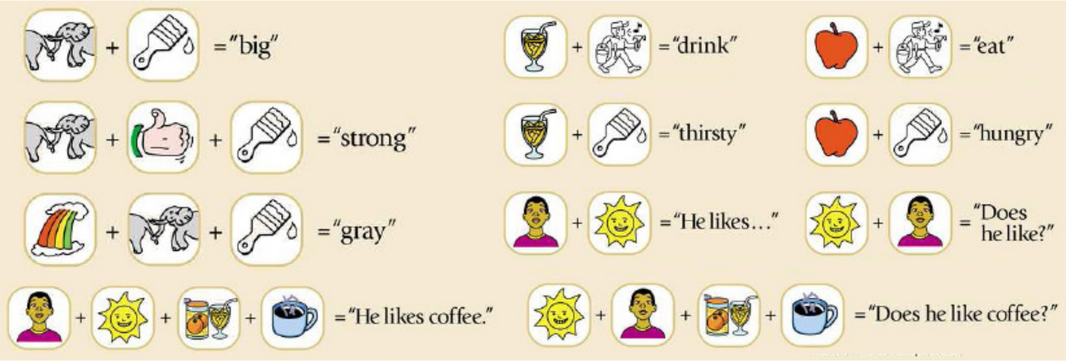


Figure 2-7. Examples of the semantic compaction method, the paint brush picture means "adjective" and the workman picture means "verbs."

2.3.2.2 Symbol System

After specifing the coding mechanism to map representations to words and phrases, the next step is to construct the symbol set. Table 2-1 summarizes various symbol systems (Fuller1992).

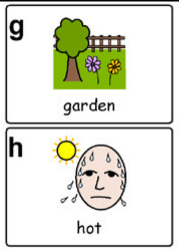
Input	motor/visual	visual	tactile/visual	auditory
Symbol System	gesture, body movement	text, pictures	objects	Morse code, speech
e.g.				

Table 2-1. Symbol systems and examples.

Since the focus of my thesis research is electronic AAC technologies, I do not include body languages and tangible objects (tactile symbols) into the following discussion. Visual representations are the dominant means to evoke meanings on AAC devices. Based on their source, the visual symbols can be categorized as stylized icons (graphic symbols created by artists or graphics designers), self-drawing pictures (by the users), user-uploaded photos, and images from the internet; based on the characteristics, visual symbols can be divided into simple linear outlines, sketches/realistic drawings, and photographs; and based on color schemes, the visual symbols can be grouped into black-and-white pictures, colored pictures, and photo-realistic pictures (Figure 2-8). A symbol system may have a mixed style of representations. Different individuals may prefer different representations; and the efficacy of various symbols changes from word to word, from person to person. The choice of a symbol system should take into consideration the users' age, gender, cultural background (Huer2000), cognition, vision, physical abilities, and literacy skills. The power of pictures in conveying meanings and the comparison across all kinds of pictures is described in *Chapter 2*.

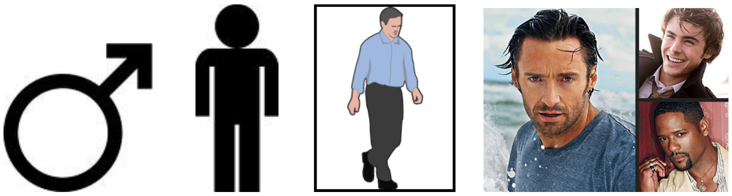
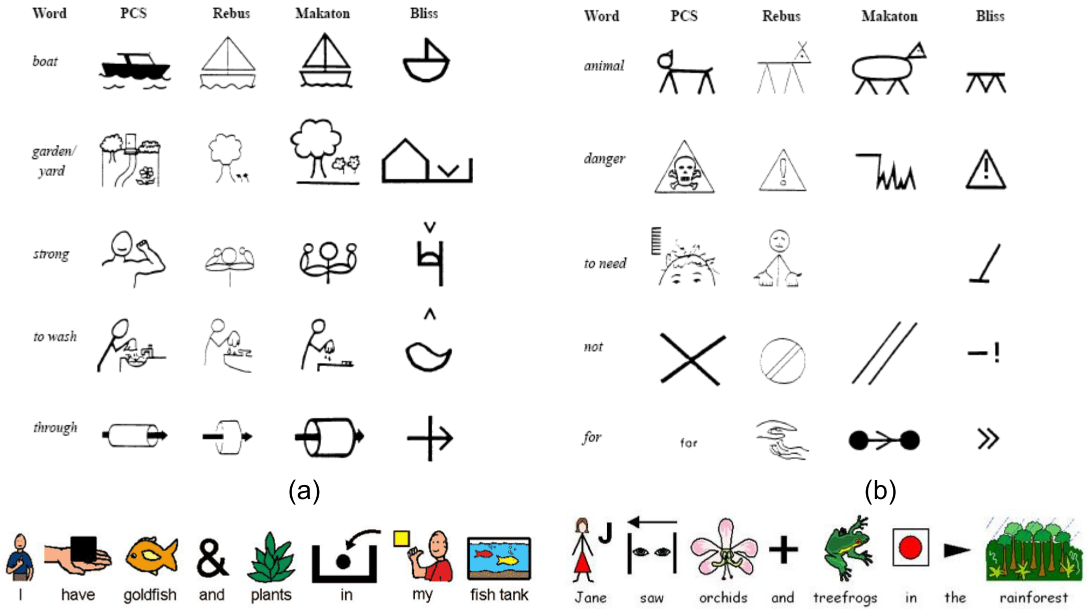


Figure 2-8. Examples of symbols for “man” in different styles.

Stylized graphic symbols, also referred to as **icons** in AAC literatures, are the conventionally used representations in AAC, since when the development of AAC systems started in early 90’s, computer graphics design and photo taking techniques were still limited. It was hard to include the AAC users into developing the symbol inventory, and thus had to rely on the specialists. A set of criteria for the icons were proposed (Mealing1995): 1) graphically clear; 2) semantically unambiguous; 3) without cultural racial or linguistic bias; 4) adaptable; 5) simple. Factors which may be added to enrich the icons include color, movement (meaning animations), and background visual cues. The effects of these factors have not yet been investigated. Some examples of graphic symbol systems developed in the early and mid 90’s are shown in Figure 2-6 and 5: Picture Communication Symbols (PCS, from Winslow), Rebus (from the Widgit Symbols Development Project), Makaton (from the Makaton Charity, originally based on Rebus), Blissymbol (from Blissymbol Communication UK), and Picto (from Pictoworld, Figure 2-6).



(a) (b)

(c) (d)

Figure 2-9. Samples of PCS, Rebus, Makatom, and Bliss icons on concrete (a) and abstract (b) words; example of sentences built with PCS symbols (c) and Rebus symbols (d).

As shown in Figure 2-9, other than Blissymbol (the original goal was to create an international written language), the visual content of the different icon sets are similar for concrete words, but varies greatly for abstract words. The icons are simple line drawings, with colors added to the later versions (see sentences in Figure 2-9).

Along with the development of graphics design techniques, the style of the icons gets more realistic and sophisticated. The change of Lingrahica icons (from Lingraphicare, Figure 2-10 (b)) is an example of this trend. On the left are the icons inherited from C-VIC (Computer-based VIsual Communication, Steele1989) which was developed from the 80's. On the right are the icons currently used in Lingraphica devices. SymbolStix is a new symbol sets that has been applied on new AAC applications (e.g. Proloquo2Go). It has a mix style of line drawings and realistic pictures (Figure 2-10 (a)).

All of the symbol systems mentioned above are using the single meaning picture coding mechanism. The sentences created from the symbol system have a picture for each of the words in it. The size of the symbol sets are huge: by 1998, PCS had around 3,000 icons, Rebus had about 950, Makaton had about 1,000, and Blissymbol had about 2,500. By the end of 2009, Lingraphica had over 3,600 icons, and SymbolStix had over 4,000 icons/drawings. The number is still growing. Minspeak (Figure 2-10 (c)) is the symbol system that is created for the semantic compaction coding scheme. The sentence constructed using Minspeak is shown in Figure 2-10 (f), in which a sequence of icons is used for each word. In 1998, Minspeak created a maximum of 128 icons for its vocabulary.

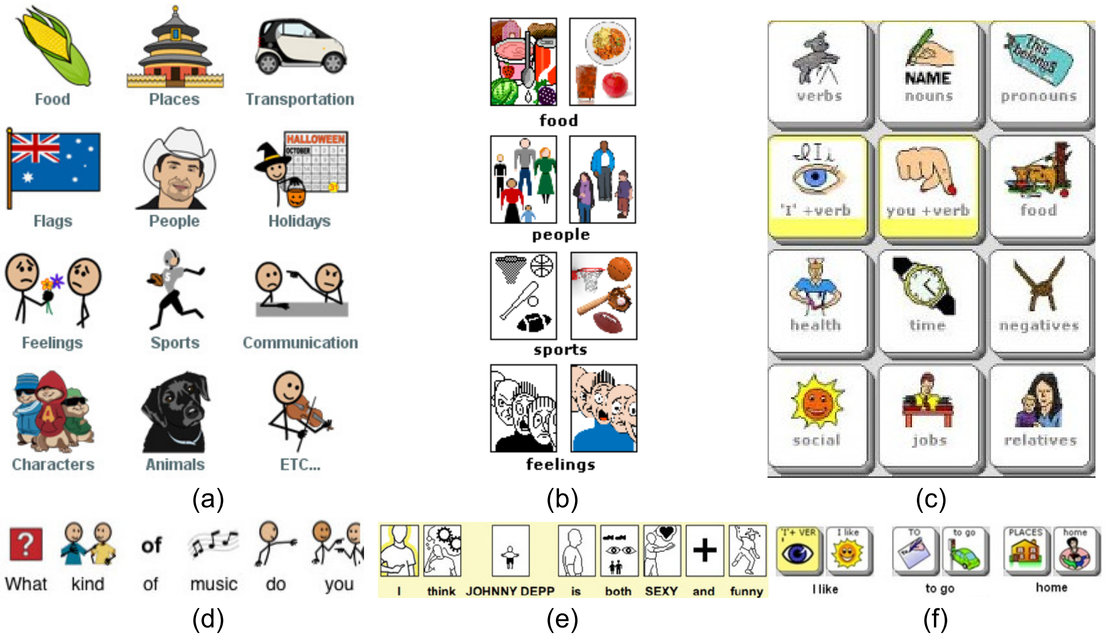
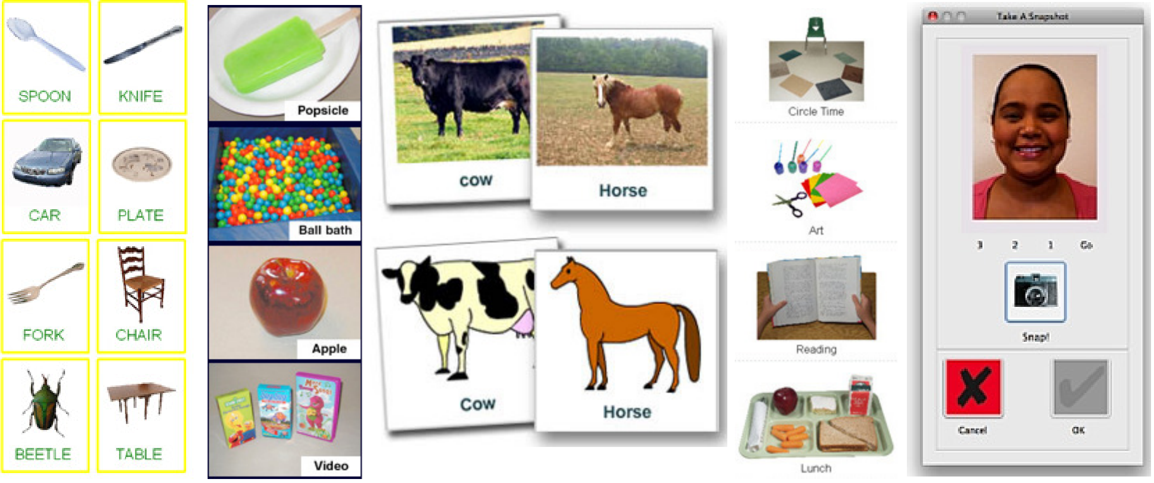


Figure 2-10. Samples of SymbolStix (a), Lingraphica (b), and Minspeak (c) icons; example of sentences built SymbolStix (d), Lingraphica (e), and Minspeak (f) icons.

Photo datasets have been added to the collection of visual representations. Here are some examples, Learning Master (LearningMaster2010, 2,700 photos, Figure 2-11 (a)), Elliecards (Elliecards2010, 250 photos, Figure 2-11 (b)), Mrs. Riley (Mrs. Riley2010, 3,000 symbols, Figure 2-11 (c)), Silver Lining Multimedia, Inc Visual Essential (SilverLining2010, 3,500 photos, Figure 2-11 (d)). Unlike the graphic symbol sets which are employed in actual AAC devices, these photo datasets are created for general language learning/rehabilitation. However, compared to the stylized symbol systems which are designed by specialists, normal people can participate in creating visual representations with photos. Lingraphica (Figure 2-11 (e)) and Mrs. Riley, for example, provide software support that allows users to make an icon/picture card with self-uploaded photos or images found on the web. This opens up the possibility of personalize the AAC vocabulary.



(a) Learning Master (b) ellicards (c) Mrs. Riley (d) Visual Essential (e) Lingraphica

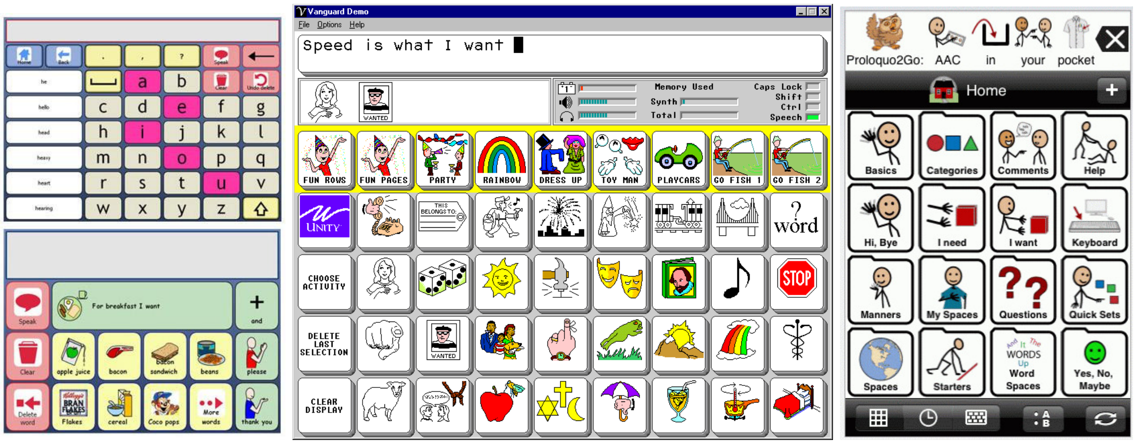
Figure 2-11. Examples of photo symbol systems.

2.3.3 AAC Vocabulary Organization

An AAC vocabulary tends to have over thousands of symbols associated with words and phrases. How to display these symbols in a structured way to facilitate efficient and effective navigation/retrieval is an essential problem (Beukelman and Mirenda2005). Three kinds of display mechanisms have been used in AAC systems: grid display, visual scene display, and hybrid display.

2.3.3.1 Grid Display

Grid display, as indicated in the name, divided the screen (also called a page in some AAC devices) in to rows and columns. Each cell contains a symbol unit. Grid display, similar to laying keys on a keyboard (Figure 2-12 (a)), is the most common organization used in both computer-based and mobile-based AAC devices, with both single meaning picture coding and semantic compaction coding (Figure 2-12 (b)(c)). In some modification of the grid display, instead of strictly lined up, items are scattered around (Figure 2-13).



(a) Keyboard and symbol array (b) PRC device (Minspeak) (c) Proloquo2Go (SymbolStix)

Figure 2-12. Examples of grid display.

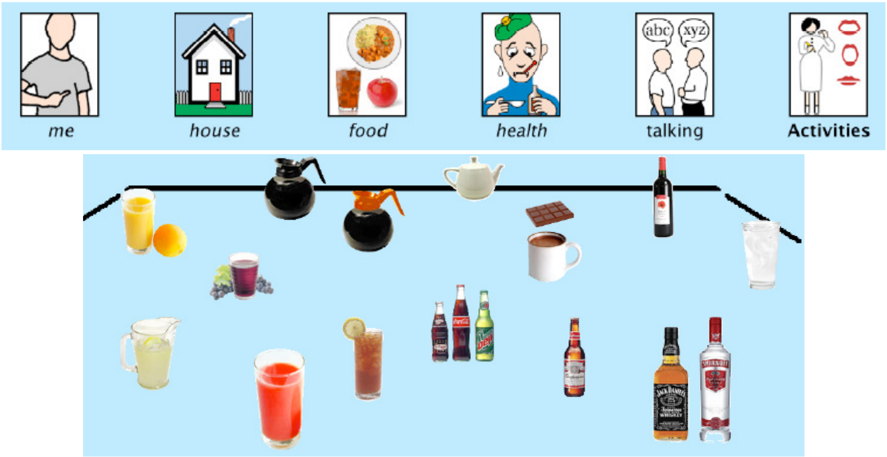


Figure 2-13. A typical grid display (top) and a scattered grid display (bottom) in Lingraphica.

The limitation of screen size determines that the big symbol vocabulary has to be divided into multiple pages, and the pages are organized hierarchically. How the symbols are grouped and where they are placed are based on how the vocabulary is selected (see the *AAC Vocabulary Selection* section). The most common strategies are 1) by **frequency**: if the words in the vocabulary are picked because of their usage frequency, the most heavily used terms appear on the main page, with pointers to go to the pages with less important terms. For example, the function words of a semantic compaction vocabulary will stay on the top page since they are necessary for each icon sequence (Figure 2-10(c) and Figure 2-12(b)). Also, the core vocabulary is usually kept on the first few pages, while the fringe vocabulary often comes late in the hierarchy.

Organization by frequency allows users to bring up the most common words quickly. 2) By **category**: if the initial vocabulary was constructed around categories, the symbols are likely to be displayed this way. The broader categories are on the top of the page hierarchy, and the further down, the terms become more specific. Figure 2-14 shows the category hierarchies in the Lingraphica vocabulary. Organization by category allows users to traverse for specific meanings easily. 3) By **activities/scenarios**: words and symbols can be grouped by their usage scenarios if the vocabulary was selected from materials about AAC users' daily activities. For example, in Proloquo2Go (Figure 2-12(c)), the scenarios on the first page include "Hi, Bye," "I want/need," "help," "questions," "Yes, No, Maybe" etc. Organization by scenarios allows users to quickly compose simple sentences for a specific setting.

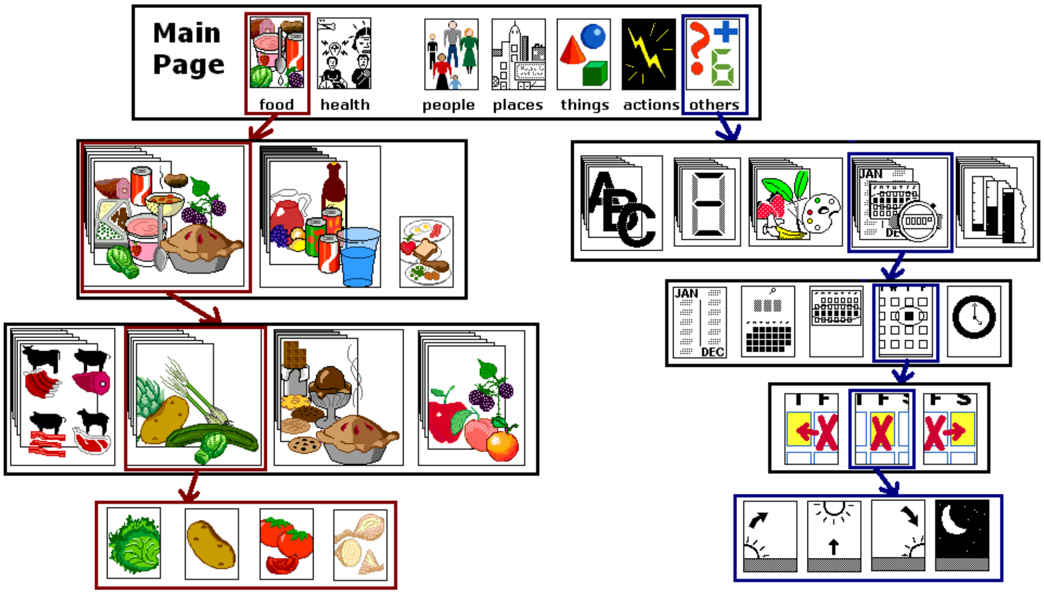
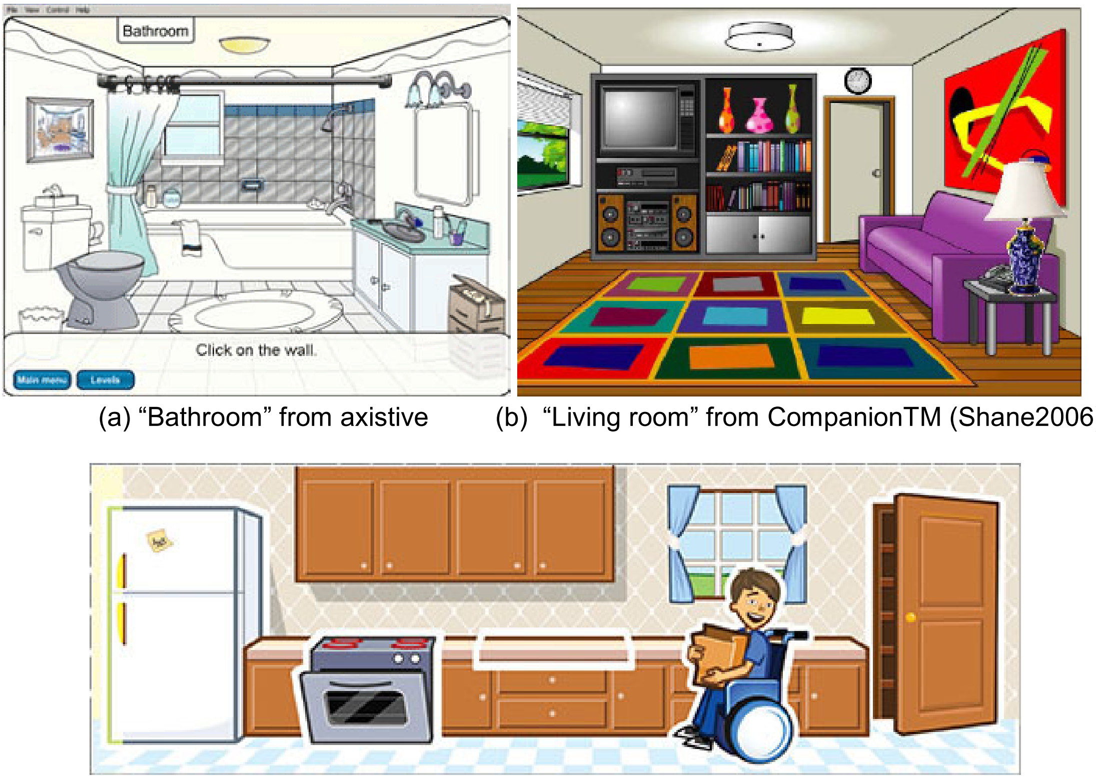


Figure 2-14. Lingraphica page hierarchy – grid display by category (food and time).

2.3.3.2 Visual Scene Display

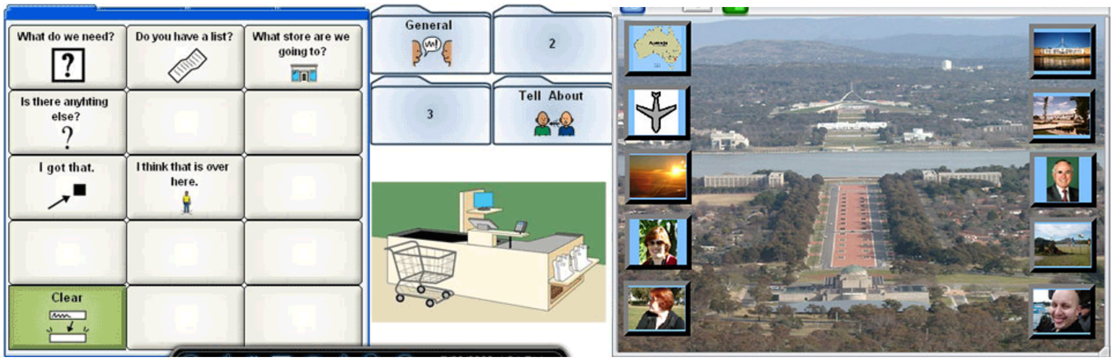
Visual scene display (or schematic scene), on the contrary to grid display, mimics real-life environments that people encounter. A visual scene display is like a photo of actual scene, a game scene, or virtual reality. It portrays people, objects, actions and events associated with a particular place (Figure 2-15), situation or experience (blackstone2004). If the AAC vocabulary was created around ecological/environmental Inventories or usage scenarios, the words in it can be arranged in a visual scene. Usually, the people and objects appearing in a scene are marked out as clickable hot spots. For examples, in Figure 2-15(c), the fridge, the stove/oven, the counter, the boy in the wheelchair, and the door are highlighted with bold white outlines, indicating that they are the symbols from the representation system. However, the hot spots are not always clearly marked (Figure 2-15(a)(b)). An AAC system can also provide words and phrases about the activities that are commonly associated with the scene. For example, as shown in Figure 2-16(a), sentences like “Do you have a list?” and “I think it’s over there.” are attached to the shopping scene. Things and events that are not shown in the static scene can be linked to it as well. Take Figure 2-16(b) as an example, the event of taking a flight and sunset accompany the main visual scene (on the left). People can create their own visual scene with personal photos (Figure 2-16(b)). It is a way to customize the AAC vocabulary.



(a) “Bathroom” from axistive (b) “Living room” from CompanionTM (Shane2006)

(c) “Kitchen” from Tangoramas

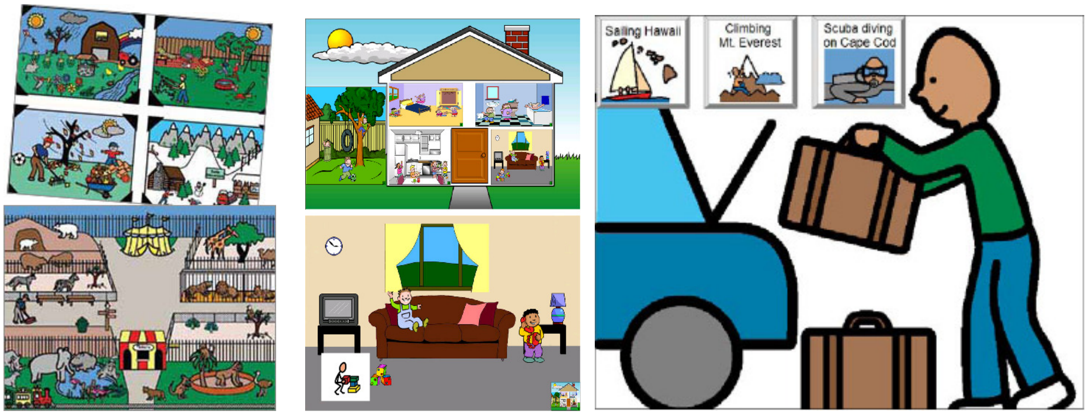
Figure 2-15. Examples of visual scene display with objects and people.



(a) visual scene with activities (b) visual scene with related events

Figure 2-16. Examples of visual scene display with activities and events (from DynaVox).

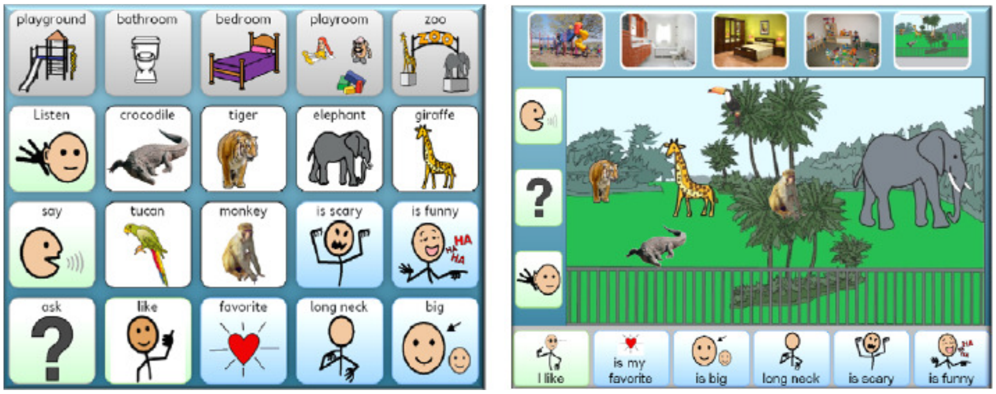
A visual scene display contains multiple pages as well. The connection between pages is different from the grid display. Different organization mechanisms allow different means to traverses among pages. The first way is to keep all the visual scenes which do not have logic relation between one another in a pool, and let the user pick out which scene he/she wants to look at (Figure 2-17(a)). In the second case, if the visual scenes are about places, the user can get to a more specific scene by navigating in a bigger scene. For example, the visual scene of a living room can be pulled out by selecting from the scene of a house (Figure 2-17(b)). In the third case, if the visual scenes depict events that are in logic or time sequence, the users can get from one view to another through a storyline or a timeline (Figure 2-17(c)).



(a) scene selection (b) scene navigation (c) scene sequences

Figure 2-17. Visual scene organization.

Previous research (Wilkinson2004) has compared the grid display and the visual scene display (Figure 2-18). From the simplicity point of view, a grid display draws clear boundaries and space between items. The layout is clear with no/little distraction from the background. People may get confused if they click on an object in the picture that they consider as an active link but nothing happens. People can systematically navigate in a grid display, row by row or column by column. On the contrary, it may be hard to distinguish the hot spots from the background in a visual scene Display. People need to explore in the picture to locate specific item, and a scene could be too crowded that people loose focus. However, a visual scene display can provide contextual support via visual cues in the background. Examples of visual cues are relations and relative positions between objects. The contextual information can help identify the meaning of the symbols, e.g. a cooking *pot* vs. a flower *pot*. Visual scene is also a more natural way for people to communicate (Woodbury1998), with lower mental demands on accessing information like categories. Research (Fallon2003) (Drager2004) has also shown that a visual scene display is easier for children to learn and use than a grid display.

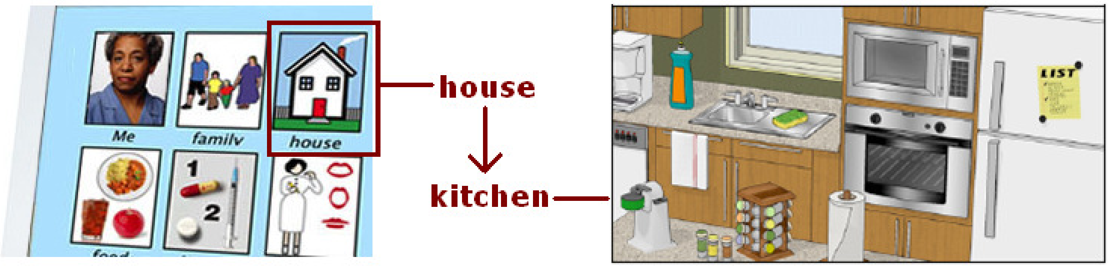


(a) grid display (b) visual scene display

Figure 2-18. Comparison between the grid and the visual scene display for "zoo."

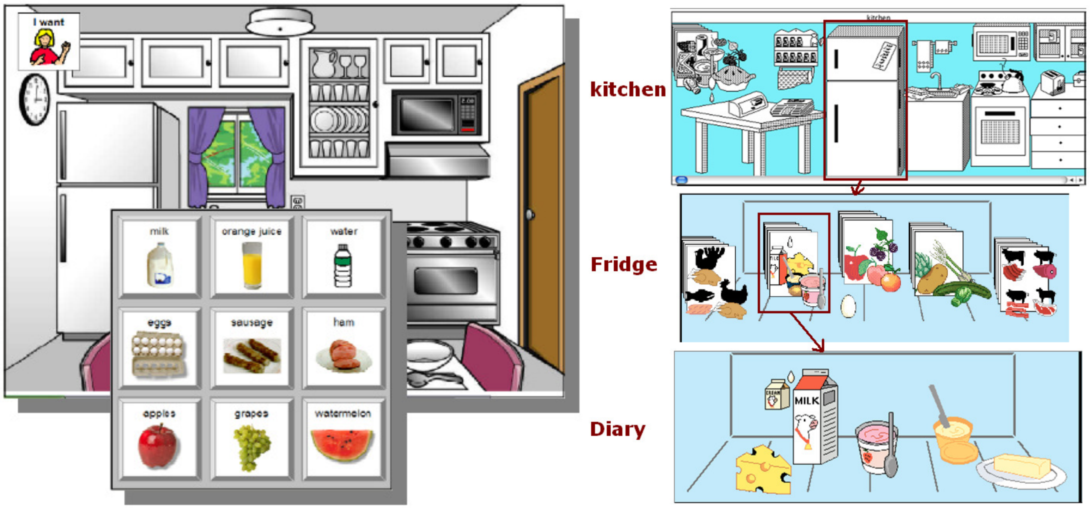
2.3.3.3 Hybrid Display

Hybrid display is the third kind of AAC vocabulary organization mechanism that combines the grid display and the visual scene display. Lingraphica for example, employs a hybrid display (Figure 2-19). In a hybrid display, the grid organization can used higher up in the hierarchy to provide structural information like category or scenario. The grid organization can also appear when many items need to be listed within limited space or when there is not much contextual information to provide (Figure 2-20(a), items in a fridge). The visual scenes are used at the middle level, putting words and symbols in context to help create an integrated sense. The grid display and the visual scene display can each take up a separate page, and a new view is open when entering a different page (Figure 2-20(b)). In some other cases, the mixed displays can appear on the same view, avoiding switching between contexts (Figure 2-20(a)).



(a) Main page grid display (b) sub-category visual scene display (kitchen)

Figure 2-19. Hybrid display - Lingraphica.



(a) popup display (b) across multiple pages

Figure 2-20. Different hybrid display mechanisms, finding “milk” in the “refrigerator.”

2.3.4 AAC Vocabulary Access, Feedback, and Output

Given an organized vocabulary/symbol system, a question that arises is how one can navigate in the structure to locate specific word. There are two main means to access a term in the vocabulary: direct selection through pointing and clicking, and scanning (especially for people with motor impairment). The direct selection can be done through a mouse or eye gaze. Scanning works in the way that the system automatically walks through each item on a page, and the user indicates a stop (e.g. by making a sound or eye blink) when the expected item is currently under scan. A grid display is normally easier to implement a scanning access (first across rows and then within a selected row) than a visual scene, in which there may not be any systematic order among items.

The AAC system needs to notify the user once a particular symbol is selected. The usual ways to provide such feedback include changing the appearance of the symbol (e.g. inverted color or magnification), playing a sound (either a touch-tone kind of sound or speaking the associated word aloud), and animation (Strauss). The feedback needs to catch people’s attention without intruding the display or delaying the reaction of the system. For example, some users complain that the playback of the speech sound is too slow and the device just sits there waiting for it to finish.

As the main goal of AAC is to facilitate communication, the main output is usually phrases or sentences in a triplet of words, symbols, and speech sounds. The idea is to let an AAC device speak for you. As people practice composing sentences via finding words in a vocabulary, it is likely that they will learn/relearn language skills. And thus, as a sub-outcome, AAC technologies have been used for speech-language pathology, language rehabilitation, and language learning.

2.3.5 Lingraphica: An Example of AAC Devices

As mentioned earlier, an AAC system is an instance of the integration of three components: a representation system, a communication medium, and a way to access and transmit messages. An AAC device is a physical package of an AAC system. Some successful cases of AAC devices are shown in Figure 2-21. Lingraphica will serve as an example to summarize all the aspects that have been discussed.

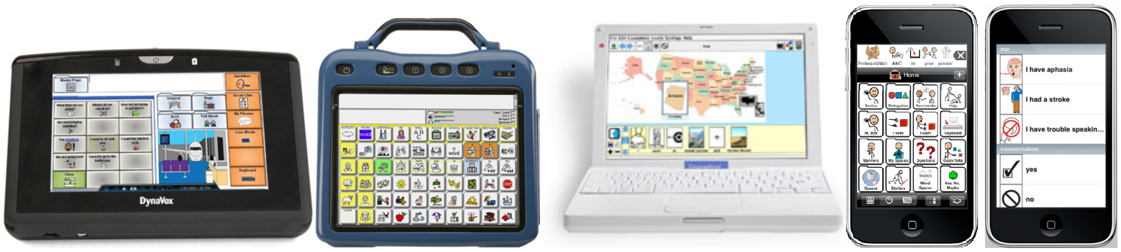


Figure 2-21. Examples of AAC devices, from left to right: DynaVox, PRC device, Lingraphica, Proloqua2Go on iPhone, and SmallTalk Lingraphica on iPhone.

Lingraphica is a commercial communication support device (a dedicated apple iBook with the Lingraphica software) for people with aphasia. Lingraphica has a core vocabulary of over 3,600 commonly used words and phrases, and features that allow users to create personal fringe vocabulary (Figure 2-11(e)) through embedded camera and web search – **vocabulary selection**. Lingraphica assigns each word/phrase an icon (**single meaning picture encoding**), but it also allows high functioning users to search for words by typing the letter sequence (**full spelling** + **word prediction**). Lingraphica has a mixed style of icons due to the currently undergoing changes between versions of the symbol system. Some of the icons are black and white simple line drawings, some are colored drawings, and some are realistic pictures and photos – **vocabulary representation**. Lingraphica employs the **hybrid display** mechanism to structure its vocabulary. Words are grouped into different pages by **category** and **visual scenes**. Users navigate in the hierarchy and select items on the display by clicking. The chosen icon enlarges and the speech sound of the assigned word/phrase is played. Multiple icons can be selected to compose longer expressions and sentences, and Lingraphica can play the synthesized speech sound for the entire sentence. Lingraphica also provides therapy excises such as spelling, grouping, and pronunciation for people with aphasia.

2.3.6 Problems with Current AAC Vocabularies

Surprisingly, from interviews with speech-language pathologists (SLPs) who work closed with people with aphasia for language rehabilitation and interviews with customer support for Lingraphica, we learned that people with aphasia seldom work with the AAC devices outside of their rehabilitation training, and almost never use the systems to help communicate to others. There are many reasons for this. First, the use of an AAC device requires training and practice, which is hard to accomplish without the help of an SLP or caregiver. Second, some of the AAC devices are too bulky to carry around, and people feel embarrassed having a machine talking for them. A lot of the people simply use the mobile device with AAC software installed as a notepad. Third, and the most importantly, people often cannot compose phrases quickly enough to carry a normal conversation. This indicates that there are severe problems in the current AAC vocabulary, which are identified and listed below.

(1) **Poor scalability** (personalization ability) of the AAC vocabulary.

The topics and words needed to talk about those topics vary greatly from person to person, especially for people with language disabilities who already have restricted access to the outside world. The ability to expand the fringe vocabulary is essential for AAC users. In order to add a word/phrase to the vocabulary, a symbol illustrating that term needs to be created as well. Some AAC devices (e.g. Lingraphica) allow users to create icons from uploaded photos. However, a user has to make the effort taking a photo, transfer it, tailor it to the right size, and then tag it with the assigned word. It is a heavy burden, even if it is the SLP or caregiver who does the work.

(2) **Low efficacy** in the **representations**.

A lot of the times, AAC users rely on the non-verbal representations to understand the words in the vocabulary. If the representations cannot actually convey the pre-assigned meaning, people have trouble finding the words they want to express. For example, in a usability study we conducted with Lingraphica, half of the people mistook the hospital icon for something else, e.g. a church or a school. Unfortunately, little work has been done verifying the efficacy of the existing symbol systems which only reflect the idea of the artists/designers.

(3) **Ambiguities** in the **representations**.

In both the single meaning picture coding and semantic compaction coding, pictures or picture sequences are mapped to individual words, not specific word senses. As discussed in the Language and Communication section, words are polysemous, meaning that a word can be associated with multiple concepts. If the representation of one concept is used to communication another, people will get confused. For example, a “letter” could mean a character or a written message. Miscommunication may occur if the picture of an envelope is used in the phrase “a missing letter” with the actual intent being a character in a word is missed out.

(4) **Low access efficiency** due to the **organization** of the AAC Vocabuary.

People have shown difficulty in navigating in the AAC Vocabularies. In some cases, it is because the huge size of the vocabulary makes traversing the hierarchies unmanageable, either one page contains too many items and required scrolling back and forth, or the hierarchy becomes so deep that people get lost or give up thinking they are on the wrong track. In the other cases, it is the display mechanisms that are problematic. Research (e.g. Hough1993) has shown that people with aphasia only retain partial or no categorization ability. One of the aphasic individual we work with cannot understand why “doctor” and “teacher” are grouped together, instead of “doctor” and “chair.” People can also easily get confused if the usage scenario or visual scene is not what they are familiar with. For example, the icon for “salt” in Lingraphica is placed on the “table” in the dining room. A lot of the users tried to search for “salt” in the kitchen, either on the counter or in cabinets/drawers. They got really frustrated when they found it not there where they think “salt” should be placed. One can create a customized visual scene with his/her own photo, but once again, it requires a lot of effort specifying each hot spot and providing tags accordingly.

To sum up, vocabulary scalability, representation efficacy, and vocabulary accessibility have become obstacles of the effective usage of AAC technologies among people with language disorders and barriers. Users of AAC systems often find the visual representations ambiguous, find the vocabulary hard to navigate due to its un-intuitive structure, and find it difficult to customize and expand the default visual vocabulary. Approaches that have been taken in my thesis research to address these issues are presented in the next chapter.

CHAPER 3 RESEARCH OVERVIEW: ONLINE COMMUNICATION SUPPORT VIA MULTIMEDIA

3.1 Research Problems

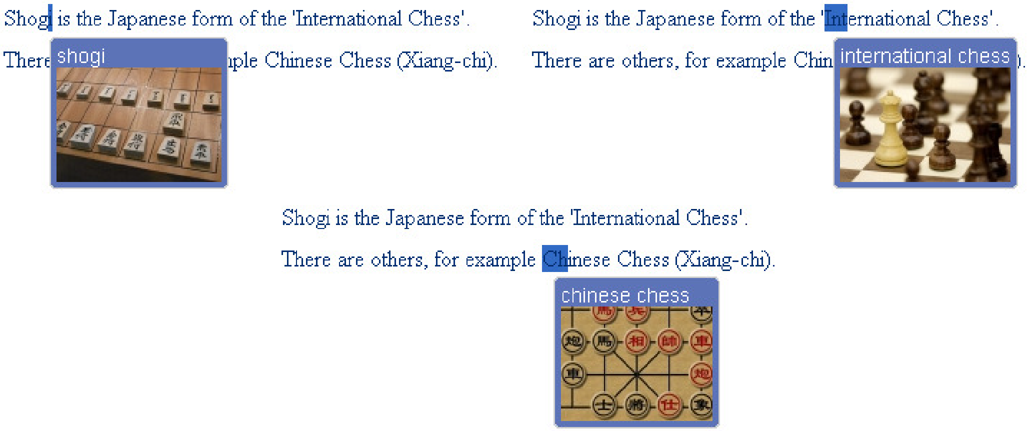
As discussed in *Chapter 2 Background Work*, natural languages are the principal way through which people communicate nowadays. The communication involves exchanging information through speech (e.g. everyday conversations) and text (e.g. books, letters, email and Internet posts). However, there are occasions where words lose their power. People with language disabilities, for example, encounter great difficulty in language comprehension and expression, and thus often fail to hold normal conversations with others and/or utilize information resources that involve text and speech. In a more general scenario, people who do not share a common language have trouble communicating with one another. People coming from a different language background often face obstacles in everyday life, and their quality of living is greatly affected.

Research has been devoted into developing Augmentative and Alternative Communication (AAC) systems that aim to facilitate communication when words are not enough, or the correct words are not known or easily accessible. However, the usage of AAC technologies has been limited due to their information evocativeness constraints and system pitfalls.

The information evocativeness constraints have several aspects. First, currently AAC systems mainly use stylized graphical icons as representations for words. Such iconic vocabulary is expensive to develop, hard to expand, and the communicative power of the representations is often not verified. Second, the representation-to-word associations in the existing AAC vocabularies are ambiguous (the words are polysemous and the contexts are ambiguous), and thus can lead to miscommunication. Third, vocabulary negation efficiency is still a bottleneck for the AAC systems due to the current ways the words and representations are organized.

The system pitfalls of the existing AAC devices involve the tradeoff between mobility and storage/display size, as well as the problems with synchronization and system update. Desktop/laptop-based AAC devices are often too bulky to use for communication outside of home. Mobile-based AAC devices are easy to carry around, but can only store limited data. Many users also struggled with the small screen and inconvenient interaction features. Current AAC systems are usually implemented as a piece of software. A lot of the times physical connection and synchronization is required for data transfer and system update.

Furthermore, existing AAC technologies mainly support communication through speech. Little work has targeted the other essential component of communication, i.e. information delivery through text. Take the Internet as an example. The Internet nowadays is one of the greatest resources of information and a convenient platform for people to communicate and network. However, this rich resource in which information is mostly in written form cannot be utilized by people with aphasia, a language disorder that can result in reading disability. People with aphasia lose the ability to recognize individual words, link them to concepts in real life, and/or understand sentences. Similarly, people (with unimpaired skills in their native language) have trouble reading Internet posts that are written in a language different from what they can speak. Language barriers are definitely a big obstruction to the utility of the Internet. In cases where multi-lingual translation is not effective, non-verbal communication can come into play (Figure 3-1).



Shogi is the Japanese form of the 'International Chess'.

There nple Chinese Chess (Xiang-chi).

Shogi is the Japanese form of the 'International Chess'.

There are others, for example Chin).

Shogi is the Japanese form of the 'International Chess'.

There are others, for example Chinese Chess (Xiang-chi).

Figure 3-1. Example of AAC assisted text comprehension.

In summary, in order to enhance the efficacy of AAC technologies and benefit a broader audience (people with language disabilities and people with language barriers) in a broader scenario (oral and verbal communication), the following research questions need to be answered:

(1) What kinds of non-verbal representations are effective, inexpensive, and easy to scale?
(2) How can information be encoded with non-verbal representations so that message can be unambiguously conveyed?
(3) How can information and illustrations be organized so that they can be efficiently accessed to support timely communication?
(4) Can AAC enhance the comprehension and usage of written information online, and how?

During my Ph.D. study, I have been exploring approaches to addressing the issues with current AAC research listed above. A variety of unconventional multimedia representations have been studied. In addition, the Internet is proposed as a new platform for hosting and utilizing AAC technologies. A novel online multimedia-enhanced communication support system was built to assist communication beyond words.

3.2 Proposed Approaches

Given the evocativeness limitation in conventionally used iconic representations, other illustration modalities should be taken into consideration. The use of web images (*Chapter 4*), videos (*Chapter 5*), and non-speech audio (*Chapter 6*) as carriers for concepts that are frequently used in everyday communication is proposed in this thesis. A large amount of these multimedia data are universally accessible through the web. They are numerous, varied, and less expensive, which diminishes the cost of expanding the AAC vocabulary. Furthermore, images, videos, and sounds open up different channels in contrast to icons for conveying information, and thus may achieve better evocativeness performance or target a different set of concepts that icons have not yet been able to effectively convey.

In order to minimize communication ambiguities and promote the access of information, I proposed to associate non-verbal representations with concepts connected in a semantic network (Figure 3-2) instead of with words organized by categories, usage frequency, or scenarios which could be polysemous. Existing AAC vocabularies have visual representations assigned to words. If the difference senses of a word are illustrated by the same picture, miscommunication could arise. For instance, the word “fly” means “travel through the air” in context like “fly to New York,” while it refers to an insect in the phrase “a fly on the trashcan.” For the word “fly,” Lingraphica (a commercial AAC device for people with aphasia) only has an icon showing an airplane and a flock of birds flying. Confusion occurs when a sentence like “I want to kill the fly (the insect)” is explained using the airplane/bird icon. Similarly, it will lead to misinterpretation if the sound of keys jingling is used to express “a key is missing” when the person means a key on a keyboard.

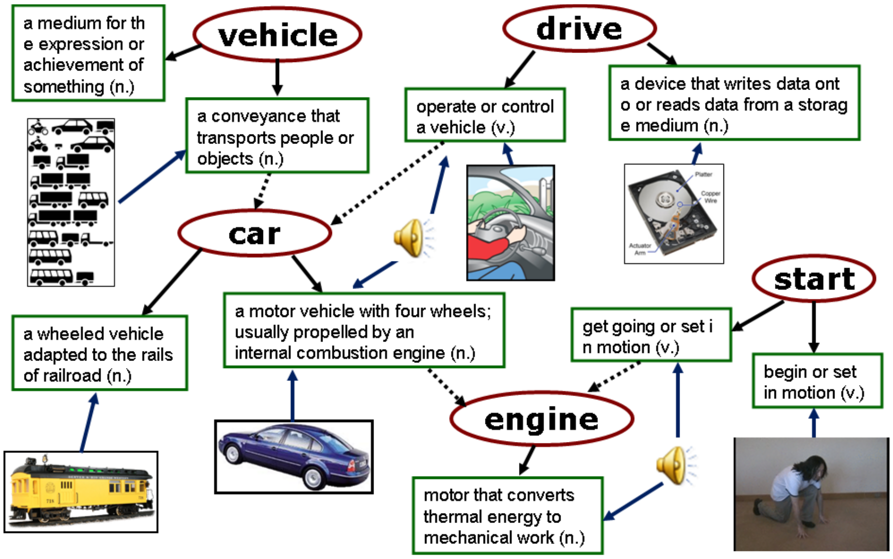
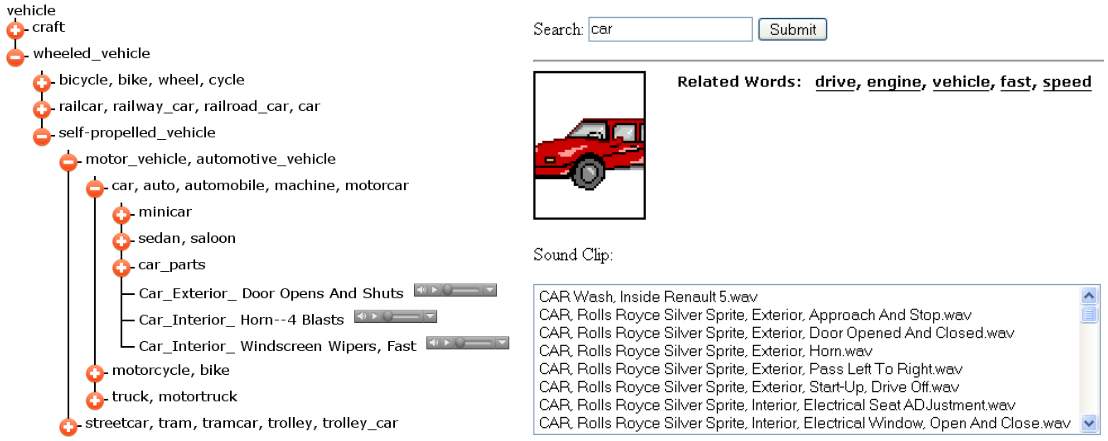


Figure 3-2. Example of the multimedia-enhanced semantic network.

On the other hand, a semantic structure provides cross-parts of speech associations between concepts which are usually missing in category/frequency-based vocabulary structures (Figure 3-2). By explicating the connections between meanings (e.g. Figure 3-3 left) and suggesting concepts that are semantically related based on the current context (e.g. Figure 3-3 right), the effort that one needs for word finding by traversing a complex/non-intuitive vocabulary is reduced.


vehicle
craft
wheeled_vehicle
bicycle, bike, wheel, cycle
railcar, railway_car, railroad_car, car
self-propelled_vehicle
motor_vehicle, automotive_vehicle
car, auto, automobile, machine, motorcar
minicar
sedan, saloon
car_parts
Car_Exterior_ Door Opens And Shuts
Car_Interior_ Horn--4 Blasts
Car_Interior_ Windscreen Wipers, Fast
motorcycle, bike
truck, motortruck
streetcar, tram, tramcar, trolley, trolley_car
Search:
car
Submit
Related Words: drive, engine, vehicle, fast, speed
Sound Clip:

CAR Wash, Inside Renault 5.wav
CAR, Rolls Royce Silver Sprite, Exterior, Approach And Stop.wav
CAR, Rolls Royce Silver Sprite, Exterior, Door Opened And Closed.wav
CAR, Rolls Royce Silver Sprite, Exterior, Horn.wav
CAR, Rolls Royce Silver Sprite, Exterior, Pass Left To Right.wav
CAR, Rolls Royce Silver Sprite, Exterior, Start-Up, Drive Off.wav
CAR, Rolls Royce Silver Sprite, Interior, Electrical Seat ADJustment.wav
CAR, Rolls Royce Silver Sprite, Interior, Electrical Window, Open And Close.wav

Figure 3-3. Example of interfaces for accessing a semantically structured multimedia dataset.

However, building such a multimedia-enhanced semantic network from scratch is a lot of work. Doing this by hand is not feasible considering the requirement of human labor and time. As an alternative, I propose a mechanism (described in *Section 4.2*) to automatically construct semantic linkage between multimedia data and concepts. The mechanism takes in large scale online multimedia label datasets (i.e. inventories of tagged images, videos, and environmental sounds) and automatically discriminates the meanings of the labels based on a combination of various semantic relatedness measures. In this way, multimedia data become the representations for the specific concepts denoted by their labels, and these concepts are connected through links defined and weighted by the semantic relatedness measures. Evaluation shows that this mechanism can achieve over 80% accuracy in label sense disambiguation, and thus can save a lot of effort on manually identifying related concepts for each multimedia representation. It greatly improves the scalability of the AAC vocabulary built upon the rich online multimedia resources.

However, one cannot take for granted that random images, videos, or sounds collected from the Internet can effectively convey intended meanings, especially given the large variation in the quality of these multimedia data. Therefore, the communicative power of the new representations needs to be verified. I proposed and conducted a series of six studies that evaluated the representation efficacy of web images, videos, and environmental sounds, in comparison to the conventionally used stylized icons and animations. The studies included controlled experiments as well as online surveys using a new web platform, i.e. Amazon Mechanical Turk. Strategies were applied to ensure the quality of results from the online surveys which are less strictly manipulated than controlled lab experiments. The studies were conducted with our target population (people with language disabilities and people with language barriers) and simulated users (at the initial stage of the research or when the target population was inaccessible).

Two aspects were examined to determine representation effectiveness: what kinds of concepts are easy/hard to illustrate and what kinds of visual/auditory cues can enhance/hinder people's perception and comprehension. A set of guidelines for designing effective concept carriers using static visual representations, dynamic visual representations and auditory representations were proposed based on the analysis and observation in the evaluation studies.

To resolve/bypass the system constraints that current AAC devices are suffering, I proposed the use of web servers as the host and a browsers extension as the interface to the multimedia enhanced vocabularies (*Chapter 7*). Such an online system allows users to access the AAC vocabulary from any device equipped with an Internet connection and appropriate web browser. No local storage of the potentially massive multimedia inventory is required, and the centralized updates of the system can be easily obtained without physical synchronization. The web browser extension also allows users to personalize the vocabulary interface (e.g. selecting preferred multimedia stimulus or preferred representation for specific concepts). Customization information can be utilized by the server for vocabulary refinement. The browser extension provides facilities that allow users to expand the vocabulary by uploading images, videos, and sounds. The new data can be shared across all users through the network.

The online multimedia AAC vocabulary can provide appropriate representations for assisting ongoing conversations, what is more, it can help to explain unfamiliar terms that people encounter when browsing web pages. It is useful for people with language barriers to look up information on the Internet. I proposed and carried out user studies in two scenarios, food and medical care, to examine if the online multimedia communication support system can enhance people's comprehension of information given in an unfamiliar language (described in *Chapter 7*). Results showed that not only non-native but also native speakers can benefit from the communication support via multimedia representations.

3.3 Contributions and Broader Impact

The contributions of this thesis research can be summarized in three categories.

(1) **Theories**. This research is the first that systematically compares four different visual stimuli plus one auditory stimulus as concept carriers for communication. I verified the communicative power of various representations, which is provides the theoretical grounds for using such stimuli in Augmentative and Alternative Communication techniques. Especially,

the exploration of videos and environmental sounds as well as the concept-level representation assignment is novel. Representation design guidelines based on empirical data will direct future research in the area.

(2) **Methodologies**. My contributions in methodologies come from two aspects: methods to automatically construct and organize a semantic multimedia dataset, and methods to evaluate such a dataset with a large number of human participants from a broad spectrum. The former refers to the algorithms designed for unsupervised multimedia label sense disambiguation. The latter refers to the use of Amazon Mechanical Turk services and the strategies to provide quality control in studies of this kind.

(3) **Applications**. The online multimedia communication support system will be the first adaptable/adaptive system that supports individuals in need of multimedia to understand and use written language available through commonly used browsers. It can support people with language disabilities or barriers in more easily retrieving information from the Internet and in communicating via the Internet. This kind of communication assistant is extremely helpful in areas that involve intensive information exchange, such as health care.

In a broader scope, besides the target population, my research can benefit an even broader audience, including children learning their first language, people learning a non-native language, and people with low literacy, by providing effective support for language comprehension and production. Although not investigated in my thesis research, it is likely that through the use of this semantically organized multimedia vocabulary, people’s mental lexicon can be enhanced or restored. Language learners can have better understanding of how concepts are related and used in different context. People going through language rehabilitation can regain the knowledge of semantic organization. Overall, my thesis research can facilitate different aspects of communication beyond words.

Chapter 3 of this thesis summarizes the targeted research problems, proposed solutions, methodologies, and contributions of my Ph.D. research. The next chapter will start with the most basic approach of non-verbal communication, i.e. through static pictures, walk through the problem space, and introduce theories and methodologies developed on the way.

CHAPTER4 STATIC VISUAL REPRESENTATIONS FOR ALL PARTS OF SPEECH

4.1 Static Visual Representations: Perception and Usage

4.1.1 A Picture is Worth a Thousand Words: Pictorial Communication

In some literature (e.g. Huggins and Entwisle1974), natural languages were called symbolic communication, meaning that the verbal system consists of symbols with assigned nominal meanings. On the contrary, iconic (or pictorial) communication refers to non-verbal communication via visual representations that convey concepts through resemblance or analogy. Even among natural written languages, many are/were inspired by or evolved from drawings. These written systems are called pictorial writing (e.g. Chinese, Figure 4-1). In pictorial writing, characters for concrete objects are simple sketches of the appearance, while abstract ideas are represented through composition of basic symbols for concrete concepts.

Besides the development of pictorial writing, ancient cultures used pictures (paintings) to depict events and scenes in everyday life, such as the hunting/fishing paintings shown in Figure 4-1. Messages like “a harvest/big victory,” “grateful,” or “praise the brave” were expressed in the paintings.

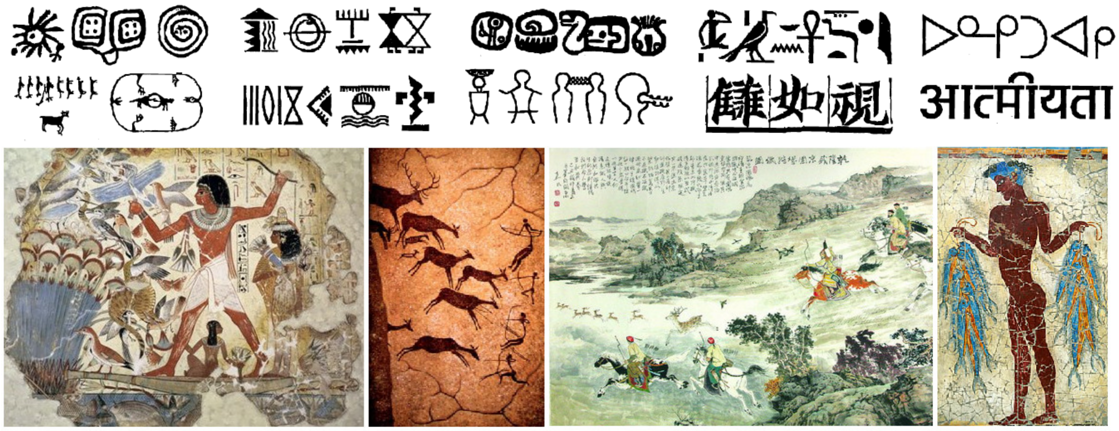


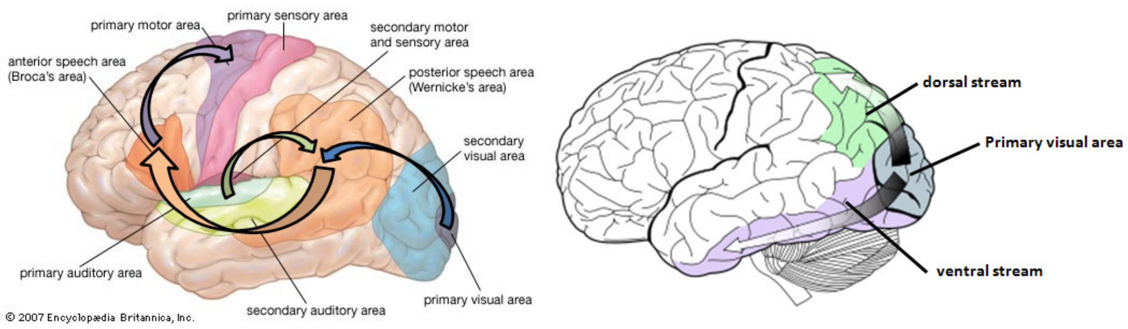
Figure 4-1. Examples of pictorial writing and ancient paintings.

Visual/pictorial communication spreads across much larger scope and appears in all sorts of everyday encounters (Figure 4-2). Graphs and figures are used to visualize data; signs and map legends are designed for traffic control, prompt information or warning of danger; magazine covers, illustrations in articles, and movie posters are summaries of the gist; advertisements mean to convey/sell an opinion or a product; various art works (photography, paintings, sculptures, architectures, fashion design, etc.) embed the thoughts and feelings of the artists/designers; cards deliver messages like greetings, wishes, and thankfulness; and comics use analogue and exaggeration to express ideas. People are very used to interpreting pictures for the hidden messages. In contrast to text, a picture encloses rich information into a glance.

Figure 4-2. Examples of pictorial communication in everyday life: graphs, signs, media and advertisements, arts, cards, maps/flags, comics, etc.

4.1.2 Process of Natural Visual Perception, Textural Perception, and Pictorial Perception

Although people can extract information out of natural visual scenes, texts, and pictorial representations, the latter part of the three perception processes within human brains are quite different (Keeler2002). All visual perception has the same start. Lights from the surroundings go through the lens of the human eye, focus on the retina (the innermost light-sensitive membrane covering the back wall of the eyeball), and form an image of the actual object or scene. The stimulus on the retina is transformed into neural impulses by the photoreceptive cells, and transmitted via the lateral geniculate nucleus to the primary and secondary visual cortex (area) of the brain (Figure 4-3(a)). Along the path, different aspects of the visual signals (e.g. color, depth, motion, etc.) are analyzed. According to the Two-Streams Hypothesis (Figure 4-3(b)), the analysis results are forwarded to the ventral stream for object recognition, and to the dorsal stream for spatial location identification. This is where the differences start to occur.



(a) (b)

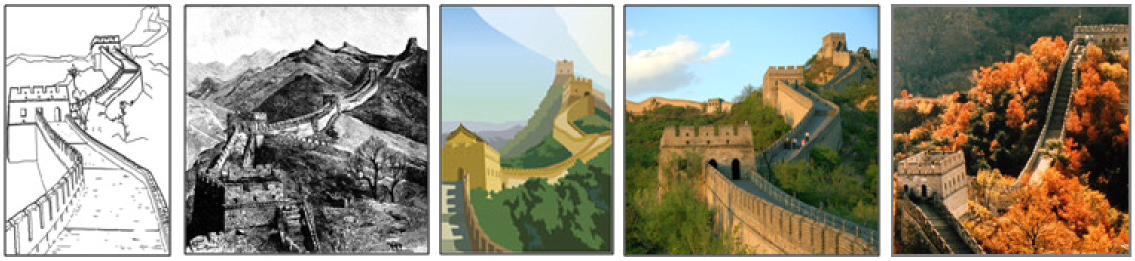
Figure 4-3. Brain areas involved in language processing (a) and visual perception (b).

Pictures are mostly presented or displayed on 2D surface. The content in a picture does not contain natural depth information, but this could be suggested through various visual cues (discussed in details in the next section). Also, information about the colors and even fine details of the shape may not be obtained from a picture. The human brain integrates various clues from the picture and reconstructs the physical, spatial and temporal image of the actual concept. However, a lot of the times, the message intended in a visual representation is not merely the objects or actions in the picture. Based on their interpretation of the visual stimuli on the retina, people use associations or even imaginations to further derive the deeper meanings from their life experiences. This step is not just a guess. Very few people will think of Saturn (planet) when they see a picture of an ant. On one hand, through common experience, people regardless of cultural background, age, and gender can identify the same concept from the same picture (e.g. international road signs). On the other hand, the same as with real scene, people can misinterpret pictures and form deceptive illusions.

Textual perception also involves associating concepts with symbolic representations, but the paths are very different from pictorial perception (Figure 4-3(a)). After the visual signals of the appearance of the words reach the primary visual cortex, they are forwarded to the Wernicke's area where the visual signals are interpreted into concepts. Wernicke's area converts speech sound into meanings as well. Later, this information is sent to Broca's area for generating the speech words in reply to what has been heard. The primary motor area is in charge of controlling the muscles to produce the voice and/or writing. The decomposition of the textual perception suggests that if Wernicke's and/or Broca's area get damaged as for people with aphasia, the ability to interprete/produce text and speech is impaired. However, as long as the brain damage does not impact the natural visual perception centers, people still retain the ability to extract concepts from visual representations.

4.1.3 Pictorial Perception

Many theories regarding pictorial perception have been proposed and studied, such as (Gibson1954)(Hochberg1962). In this thesis, I divide pictures into two generic categories: **icons** (Figure 4-4(a)~(c)) and **images** (Figure 4-4(d)(e)). Icons are a type of practical, minimalist art (e.g. line drawings and signs), while images are realistic photos. Information embedded in the pictures that helps people identify the content includes shape, depth, and color.



(a) (b) (c) (d) (e)

Figure 4-4. Comparison of simple, realistic, colored line drawing and photo of the Great Wall.

Contours formed by the lines in the picture are the basis of perception of shapes. Contours separate different surface areas as well as the object in the foreground and the background. Research (Rubin1958) has shown that lines keep and depict the most essential details of sizes and shapes. Further more, contour deformation (in size and shape) based on projection serves as visual depth cue that can help recover the spatial relationship (Hochberg1962). It suggests that simple line drawings are sufficient to convey basic objects and scenes (Figure 4-4(a)).

More sophisticated line drawings add texture and shading information to surfaces separated by contours (Figure 4-4(b)). Textures reflect the material and surface conditions. The perspective distortion of textures indicates the change of shape curvatures and surface layout (Gibson1950), as well as the graduate adjustment of position. Shadows suggest the size, position, and direction of the lighting sources. Similar to texture, shadows are cues to the transformation of shapes and the relative position of objects.

Colors further enrich the texture information. Even metals could be of different colors (e.g. red for copper, black for iron, yellow for brass). Sometimes the color could be the critical feature for making fine distinctions, such as lemon and lime. In Figure 4-4(d) and (e), the differences in colors of the foliage indicate (d) was taken in the summer or spring while (e) was taken in the fall. In colored drawings, the colors sometimes reflect the reality, and sometimes with exaggeration for the purpose of attracting attention or emphasizing certain messages. People often associate colors with particular situations or emotions. For example, red may indicate festival and happiness (in China), danger, violence, hot, and passion. Research (Pettersson1981) has shown that people in general prefer color pictures to black-white pictures of the same content. The assistance of color, texture, and visible discontinuities allows observers to identify basic features of surface like shape, slant, pigment, illumination, and pattern without training or annotations (Kjelldahl2003).

4.1.4 Comparison of Image and Icon Perception and Factors of Impact

Compared to images, icons only provide the most essential visual cues. There is less distraction in the background, and viewers can focus on the key message represented in the foreground. Icons can utilize symbols that are recognized universally (e.g. question mark and heart shape) to deliver ideas as common knowledge. Icons can also use tricks like exaggeration and direction indicators like arrows to help direct people's attention. These are all lacking in realistic photos. On the contrary, at the risk of providing too much specificity, images offer more pictorial cues than icons. Image attributes such as color, contrast, segmentation, and interaction further influence the perception of size and depth of an object. Images realistically present objects and events as they appear in the reality, and people can make interpretation based on everyday experience. No learning and memorizing are required as for certain abstract icons (e.g. some traffic signs).

There are several factors that affect people's perception of line drawings and figures. Although people from a wide variety of cultures recognize the same objects from unambiguous pictures and even detect ambiguity in similar manners, different cultures may lead to different interpretation of ambiguous pictures (Kennedy1974)(Pettersson1982). Cultural and living environment differences may influence people's perception and preferences of sizes (Turnbull1961), shapes (Lanners1973), and colors (Kurten1971) (Eysenck1959).

Aging, more specifically age-related effects like vision and cognition degeneration is another factor that can impact people's pictorial perception (Faubert2002). As people age, their sensitivity of illumination decreases, their color perception is blunted, their motor sensitivity is impaired, and they acquire loss of perception in symmetry, size, depth, and spatial frequency. These suggest that visual representations with less crowded information and bigger contrast in color and contour are easier for elderly people to perceive. Also, the differences in life experience may result in different perception of visual stimuli.

4.1.5 Pictures (Images) as a Language

In additional to the pictorial communication mentioned earlier, pictures have been used to facilitate information exchange in the absence or insufficiency of natural languages. The ubiquity guarantees that pictures can be abstracted to refer to specific entities or places; it is possible to form linguistic systems around the use of pictures. Compare to natural languages, it is less likely that learning is required for interpreting pictures; experiments revealed that infants

(Hochberg1962) appear to understand pictures without training. As pointed out earlier, unlike written words, cultural differences do not prevent people from identifying distinctive images (photographs) and icons (line drawings) in a similar way (Nadel1937). PictNet (Takasaki2006) for instance, is a pictogram communication system across cultural and linguistic barriers – between school children in Japan and Kenya. Even with an extremely terse vocabulary, these children were able to forge relationships. Pictures have also shown their power in assisting language education (Wright1990)(Carney and Levin2002).

Since pictures are an alternative method of representing a concept with often lower memory or cognitive load compared to spoken or written words (Mayer1998), pictorial representations of concepts have been widely used in the design of assistive devices for people with cognitive disabilities. Research shows that individuals with aphasia in particular retain abilities such as identification, sequencing, generalization, and meaning association, which could be used to improve communication via visual prompts (Thorburn et al.1995). Iconic languages created for Augmentative and Alternative Communication (AAC) which are mentioned in the *Background Work* section are examples of the use of icons as supplement to natural languages.

Icons are the conventional visual representations used in AAC vocabularies, however, research (Danielsson2001) suggests that introducing images as a language can lead to significant improvements to learning acquisition of people with cognitive disabilities as well as overcome the difficulty that iconic systems have in dealing with highly personalized, special areas of interest as well as categories rich with unique words. One aspect of his work showed how a person with communication impairment took advantage of digital images for rehabilitation and for improving her quality of life through better communication. The author documented that the participant in her study took photos of things that she had experienced and wanted to remember or talk about. His conjecture is that digital photographs can be used as memory support and as communication and rehabilitation tools.

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.

4.1.6 Extending AAC Vocabulary with Web Images

Although many systems for assisting individuals with aphasia have introduced photos as a supplement to their icon-based vocabulary (e.g. (Boyd-Graber et al. 2006), (Van de Sandt-

Koenderman2005)), 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 might make selecting the most appropriate image easier (or harder).

ImageNet (Deng2009) is an image database that has organized tens of millions of web images in WordNet hierarchies. On average, 1,000 images are assigned to each of the selected 100,000 synonym sets (only nouns) in WordNet. ImageNet provides a quick and free access to numerous web images that have a precise association with concepts. This kind of semantic image collection could also be used for content in an aphasia support system database. However, ImageNet currently aims to provide a dataset for computer visual (e.g. object recognition). It does not cover common concepts in different parts of speech that are used in everyday communication.

There are other resources can be utilized to construct a rich image augmented AAC vocabulary. For example, online image repositories such as Flickr (flickr2010) and Picasa (picasa2010) 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 (Von Ahn2004) might be better candidates for the purpose of building an AAC vocabulary, especially since the game ensures that each image has elicited the associated word multiple times from different people.

To sum up, static pictorial representations have been widely used in communication in aspects in everyday life. Iconic representations particularly are incorporated into Augmentative and Alternative Communication systems to support language comprehension and production for people with language disabilities. Research has shown that photo-realistic images can facilitate communication with aphasic individuals as well. Given the abundant resource of web images which are cheap, numerous, and varied, images are potential representation candidates for a scalable AAC vocabulary. In the next section, my thesis research on approaches to automatically construct an AAC vocabulary from online image label dataset is introduced. The communicative power of images is evaluated in comparison to the stylized icons used in popular AAC devices. Guidelines for better creation/selection of static visual representations for all parts of speech will be proposed.

4.2 Automatic AAC Vocabulary Construction from Image Label Dataset

As discussed in the previous section, 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. Internet users from all over the world are sharing images over the web, and thus, an AAC vocabulary based on these images does not need to depend on individual developers/designers. In addition, a lot of the public domain images are free. It suggests that incorporating web images into representations for concepts may greatly enhance the scalability of the AAC vocabularies.

However, compared to the icons which were carefully crafted to support communication, web images have inherent shortcomings. First, their complex content may evoke several meanings, and different people might interpret them differently. For instance, in a survey, half of the participants named Figure 4-5(a) as “woman/lady” and the other half considered that it represents a “hat.” Second, their size, quality and resolution vary significantly (Figure 4-5(b) for “bird”). Third, relatively few web images are labeled unambiguously. For example, the image in Figure 4-5(c) is tagged as “scenery, park, car.” There is no clue whether the word “park” in the tags means “(a large area of land preserved in its natural state as public property” or “place temporarily as park the car.” It suggests that in order to convey precise meanings for communication, web images should be assigned to specific concepts instead of polysemous words. Also, guidelines should be given to direct the selection of good image representations from a huge varied pool.

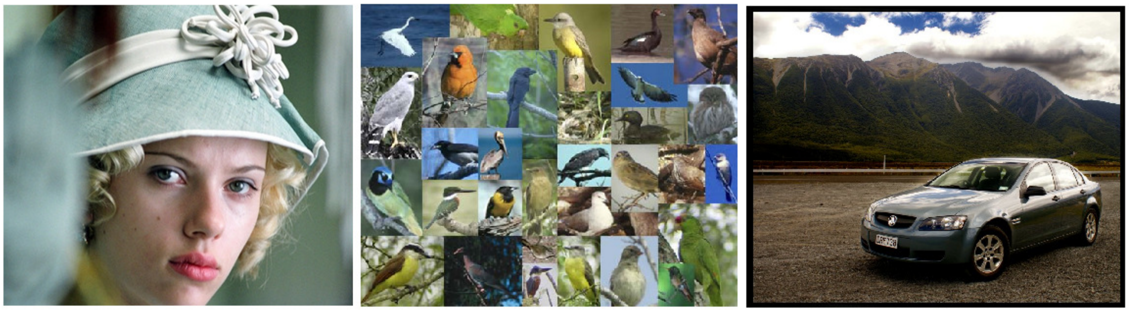

(a) (b) (c)

Figure 4-5. Examples of problems with web images.

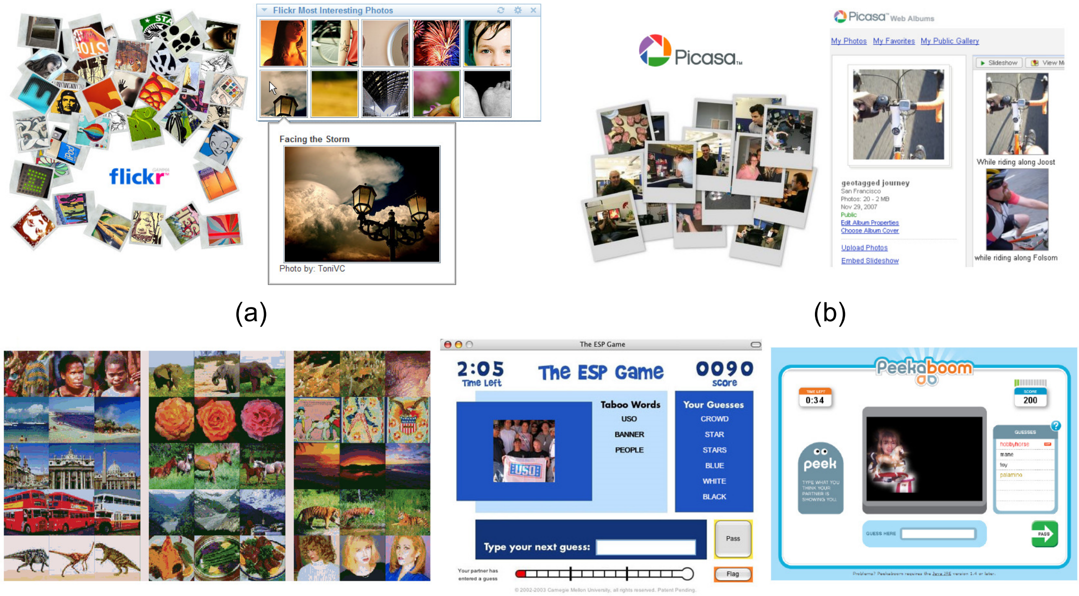
In this section, I will introduce new algorithms proposed in my thesis research that automatically generates semantic network from web image label dataset, i.e. **unsupervised image label**

disambiguation. In the next section, I will present the evaluation of the communicative power of web images compared to conventional stylized icons. Guidelines of static pictorial representation design and selection are generated from the analysis.

4.2.1 Image Label Datasets

Several image label databases have been mentioned in *Section 4.1*. For the purpose of disambiguating the meanings of the labels, my thesis research looks into only multi-label image database (excluding the single-label datasets in computer vision). Based on the source of the image captions, the image datasets can be divided into three categories: user-uploaded dataset, professional dataset, and third party-tagged dataset.

User-uploaded image datasets include online photo galleries that allow Internet users to upload, edit, tag, and share personal photos, such as flickr.com (Figure 4-6(a)), Picasa.com (Figure 4-6(b)), photo.net (photonet2010) and Windows Live Photos. The captions for each photo are usually description of the scene generated by the owners. Another kind of user-uploaded image dataset is photojournalism (e.g. panos pictures 2010 and AP images 2010). Photojournalism, journalism that presents a story primarily through the use of pictures, provides detailed descriptions around the main theme of a photo. The key people, places, objects, and events are usually highlighted in the caption (Figure 4-7(a)(b)). Images from both only photo gallery and photojournalism are more varied in terms of content and quality as well as the descriptiveness of the labels. They are not designed for the purpose of evoking specific meanings. Copyright is another issue that concerns the use of user-uploaded image datasets.



(a) (b)

(c) (d) (e)

Figure 4-6. Examples of image label database.

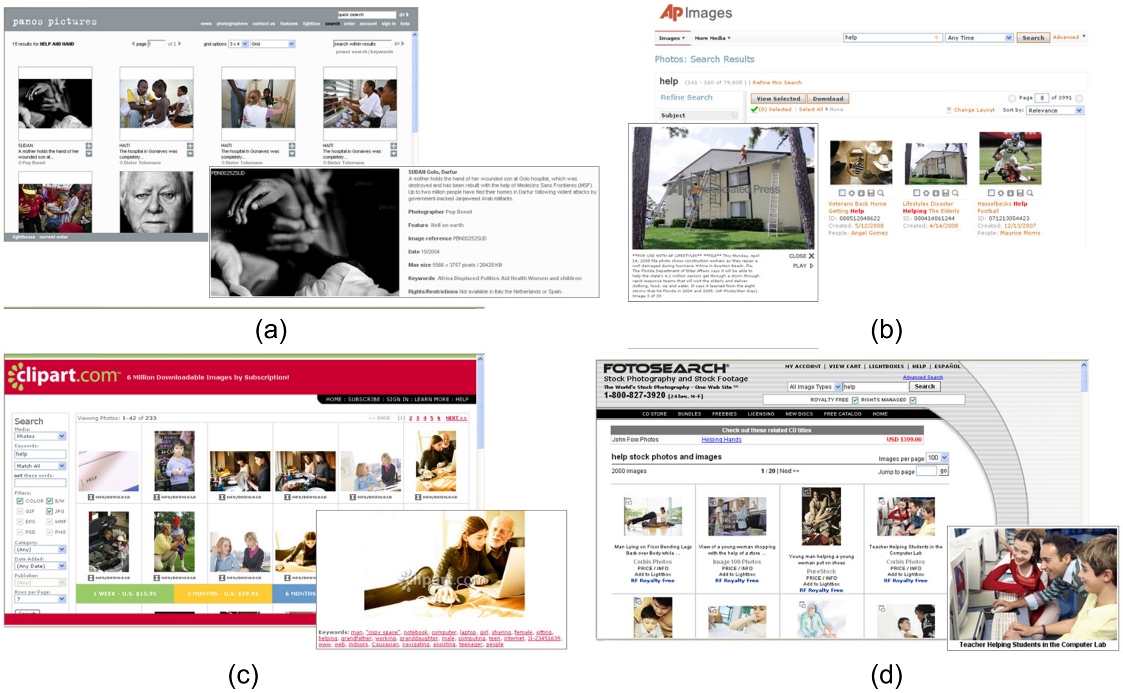


Figure 4-7. Examples of photojournalism and stock photo website; search results for “help”.

Professional image label datasets such as Corel Image Database (Corel2007) and Microsoft Research Cambridge (MSR) Image Database (MSRC2010) provide high quality images (digital photographs) organized in classes like “car,” “buildings,” “flowers,” “Africa,” etc. The Corel 10,000 Image Database (Figure 4-6(c)), for instance, has 100 image categories with 100 images each. The MSR Cambridge Database contains 4323 images divided into 23 classes. The images in professional dataset are selected by researchers and thus are more consistent in qualities. However, these datasets are mainly designed for computer vision research, such as object recognition and scene annotation. As a result, the professional databases do not cover sufficient concepts (nouns, verbs, adjectives, adverbs, etc.) to support everyday communication. Furthermore, the labels for the images in these databases are more likely to only describe objects appearing in the visual content. The labels are either too generic (meaning only the main scene/object is annotated, e.g. “Africa, people”) or too detailed (meaning everything in the image is listed, e.g. “sky, cloud, tree, grass, ground, path, bird, sun). Weak labels do not provide enough information for determine what, whereas too detailed labels lose the focus of the main object/event depicted in the images.

There is another kind of professional image label databases, i.e. stock photograph websites like clipart.com (clipart2010) and fotosearch.com (fotosearch2010). These stock picture inventories contain millions of high quality pictures by professional designers and photographers. They are great resources for interface development and artistic design. There is usually a subscription fee required for access of the images. Similar to the research-oriented professional dataset, labels assigned to each photo in the stock could be very fine-grained (Figure 4-7(c)) or very coarse (Figure 4-7 (d)).

Third party-tagged dataset refers to the kind of image inventory that has a different set of people from those select/create the images to generate image descriptions. The labels actually reflect to a certain extent people's understanding and interpretation of the image content. Two examples are the ESP Game and the Peekaboom Game datasets. The ESP Game (now Google Image Labeler) Dataset (Von Ahn and Dabbish, 2004) contains a large quantity of web images and human labels produced via an online game. In the ESP Game (Figure 4-6(d)), two randomly paired players who see the same image try to agree on a word. For example, an image of a glass of hard liquor is assigned to the labels "full, shot, alcohol, clear, drink, glass, beverage." The Peekaboom Game (Figure 4-6(e), Von Ahn et al., 2006) is the successor of the ESP Game, which asks players to specify the regions in the image that are associated with the given label (from the ESP Game). Unlike in the ESP Game, labels used in the Peekaboom Game are cleaner, and their relevance to the image can be evaluated based on the region specified and the players' performance. In my thesis research, the Peekaboom Dataset (over 57,800 images) was used. For each image, all the labels together form the context for sense disambiguation purposes.

One disadvantage of third party labels compared to self-generated labels is the noise contained in the naming process. First, random objects may appear in a picture and thus be included in the labels. For example, an image is labeled "kids, group, dance, park, tree, chief, Indian, children, sky." The words "chief" and "Indian" do not typically co-occur with "park (a recreation area)," but in this particular image, there happens to be a park staff member dressed as a Native American chief performing for a group of children. Second, texts in the images are often included in the labels as well. For example, the word "green" is one of the labels for an image with a small street sign "Green St." at the corner. Third, things that are not closely related to the main theme of the image are included in the labels as well. For instance, almost every image with an outdoor scene has the label "sky." Fourth, the Peekaboom labels for example are not stemmed, and thus another layer of ambiguity is added. For example, given only the labels "bridge, building," it is hard to tell if the image is of people constructing (building) a new bridge, or a skyline with buildings and a bridge. In the experiment, all labels for an image are used in their original, unstemmed form to form the context for the word disambiguation (WSD) algorithm.

4.2.2 Label Sense Disambiguation Algorithms

In the Peekaboom image label dataset, between two to twenty words from the human labels were associated to each image. However, given the polysemy of words, which concepts are considered relevant to the images remains unknown. It requires further work, i.e. disambiguating the senses of the label words given the image scenes. Since the image label datasets often have hundreds to tens of thousands of pictures with multiple tag,, assigning meanings to each word by hand was time consuming and labour intensive. From a broader point of view, any manual approach is of poor scalability – hard for the designers and users to add new multimedia representations for concepts that are missing in the vocabulary.

Therefore, I explored a new approach for generating a vocabulary with concept to image associations, i.e., conducting word sense disambiguation (WSD) techniques used in Natural Language Processing on the image label datasets. For example, the labels “girl, wear, hat, beautiful” for the image in Figure 4-5(a) are assigned to synsets “girl: a young woman,” “wear: be dressed in,” “hat: headdress that protects the head from bad weather; has shaped crown and usually a brim,” and “beautiful: delighting the senses or exciting intellectual or emotional admiration” via WSD. It means the image of a girl in a hat can be used to depict those concepts. This approach is viable because the words in the image labels were shown to evoke one another based on the visual content, and their meanings can be identified by considering all the tags generated for a given image as a context. With the availability of large sound/image label datasets, the vocabulary created from WSD can be easily expanded.

A variety of WSD methods (e.g. knowledge-based methods such as (Lesk, 1986), unsupervised me-thods (Lin, 1997), semi-supervised methods (Hearst, 1991) (Yarowsky, 1995), and supervised methods (Novischi et al., 2007)) were developed and evaluated with corpus data and other text documents like webpages. Compared to the text data that WSD methods work with, labels for images have unique characteristics. The labels are a bag of words related to the visual content; there are no syntactic or part of speech information, nor are the words necessarily contextual neighbors. Furthermore, few annotated image label datasets are available, making it hard to apply supervised or semi-supervised WSD methods.

To efficiently and effectively create a disambiguated audio-augmented vocabulary, two goals need to be achieved. First, optimize the accuracy of the WSD algorithm to minimize the work required for manual checking and correction afterwards. Second, ensure that the parts of speech information that is missing from the “bag of words” format is reconstructed. I targeted the first goal by proposing an unsupervised sense disambiguation algorithm combining a variety of semantic relatedness measures. I chose an unsupervised method because of the lack of a large manually annotated gold standard. The measure-combined voting algorithm presented below draws advantages from different semantic relatedness measures and has them vote for the best-fitting sense to assign to a label. Evaluation shows that the voting algorithm exceeds WSD with each individual measure significantly.

To approach the second goal, I explored the use of a new semantic relatedness measure called evocation (Boyd-Graber et al., 2006) (Nikolova et al., 2009) in disambiguation of image labels.

Evocation, as defined here, measures human judgements of relatedness between a directed concepts pair. It provides cross parts of speech evocativeness information which is lacking in most of the knowledge-based semantic relatedness measures. Evaluation results showed that the performance of WSD with evocation is no worse than most of the relatedness measures that we applied, despite the relatively small size of the current evocation dataset.

Nine semantic relatedness measures were used in the measure-combined voting WSD algorithm: (1) WordNet path based measures: “path,” “wup” (Wu and Palmer1994), “lch” (Leacock and Chodorow1998); (2) information and context based measures: “res” (Resnik1995), “lin” (Lin1997), “jcn” (Jiang and Conrath1997); and (3) WordNet definition based measures: “lesk” (Banerjee and Pedersen2002), “vector” and “vector_pairs” (Patwardhan and Pedersen2006). The computation of the relatedness scores using measures listed above were carried out by codes from the WordNet::Similarity (Pedersen et al., 2004) and WordNet::SenseRelate projects (Pedersen and Kolhatkar, 2009). In contrast to WordNet::SenseRelated, which employs only one similarity measure in the WSD process, this paper proposes a strategy of having several semantic relatedness measures vote for the best synset for each word. The voting algorithm intends to improve WSD performance by combining conclusions from various measures to eliminate a false result. Since there is no syntax among the words generated for a image, they should all be considered for WSD. Thus, the width of the context window is the total number of words in the image labels.

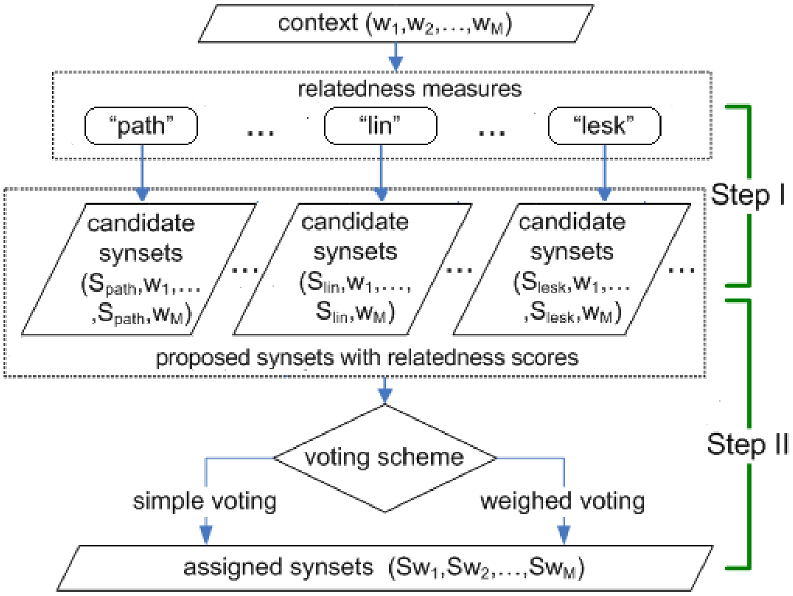

context (w_1,w_2,…,w_M)
relatedness measures
“path”
…
“lin”
…
“lesk”
Step I
candidate synsets (S_{path},w_1,… ,S_{path},w_M)
…
candidate synsets (S_{lin},w_1,…, S_{lin},w_M)
…
candidate synsets (S_{lesk},w_1,… ,S_{lesk},w_M)
…
proposed synsets with relatedness scores
voting scheme
Step II
simple voting
weighed voting
assigned synsets (Sw_1,Sw_2,…,Sw_M)

Figure 4-8. Measure-Combined Voting Algorithm.

Figure 4-8 shows the overall process of the measure-combined voting algorithm for disambiguating image labels. After the context for WSD is generated, the process is divided into two steps. In Step I, the relatedness scores of each sense of a word based on the context is computed by each measure separately. Step II combines results from all measures and generates the disambiguated synsets for all words in the image labels. Evocation did not participate in Step II.

Step I: Given the context of M words (w_1, …, w_M), and K relatedness measures (k = 1, …, K), the task is to assign each word w_j (j = 1, …, M) to the synset $s_{x,wj}$ that is the most appropriate within the context. Here, the word w_j has N_j synsets, denoted as $s_{n,wj}$ (n = 1, …, N_j). Step I is to calculate the relatedness score for each synset of each word in the context.

$$score_k(s_{i,w_j}) = \sum_{m=1,\ldots,M}^{m\neq j} \max_{n=1,\ldots,N_m}(measure_k(s_{i,w_j}, s_{n,w_m}))$$

The evocation score between two sysnets s_a, s_b is the maximum of the directed evocation ratings.

$$\underset{evocation}{score}(s_a, s_b) = \max(evocation(s_a, s_b), evocation(s_b, s_a))$$

$$\underset{evocation}{score}(s_{i,w_j}) = \sum_{m=1,\ldots,M}^{m\neq j} \max_{n=1,\ldots,N_m}(\underset{evocation}{score}(s_{i,w_j}, s_{n,w_m}))$$

The synset that evocation assigns to word j is the one with the highest score.

$$s_{w_j} = s_{x,w_j}, if \ \ \underset{evocation}{score}(s_{x,w_j}) = \max_{i=1,\ldots,N_j}(\underset{evocation}{score}(s_{i,w_j}))$$

Step II: Three voting schemes were tested, including unweighted simple votes, weighted votes among top candidates, and weighted votes among all synsets.

1) Unweighted Simple Votes

Synset $s_{n,wj}$ of word w_j gets a vote from relatedness measure k if its $score_k$ is the maximum among all the synsets for w_j, and it becomes the candidate synset for w_j elected by measure k ($C_{k,wj}$):

$$vote_k(s_{x,w_j}) = \begin{cases} 1, if \quad score_k(s_{x,w_j}) = \max\limits_{i=1,\ldots,N_j}(score_k(s_{i,w_j})) \\ 0, else \end{cases}$$

$$candidate_k(s_{w_j}) = s_{x,w_j}, \qquad if\ vote_k(s_{x,w_j}) = 1$$

The candidate list for word w_j (candidates(Sw_j)) is the union of all candidate synsets elected by individual relatedness measures.

$$candidates(s_{w_j}) = \underset{k=1,\ldots,K}{union}(candidate_k(s_{w_j}))$$

For each candidate in the list, the votes from all measures are calculated. The one receiving the most votes becomes the proposed synset for w_j.

$$voteCount(s_{i,w_j}) = \sum_{k=1}^{K} vote_k(s_{i,w_j})$$

$$s_{w_j} = s_{x,w_j}, if$$

$$voteCount(s_{x,w_j}) = \max_{s_{i,w_j} \in candidates(s_{w_j})} (voteCount(s_{i,w_j}))$$

Weighted Votes among Top Candidates

The weighted voting scheme avoids the situation that the false results win by a very small margin. The weight under relatedness measure k for $s_{i,wj}$ is calculated as the relative score to the maximum $score_k$ among all synsets for word w_j. It suggests how big of a difference in relatedness score of any given synset is to the highest score among all the possible synsets for the target word.

$$weight_k(s_{x,w_j}) = score_k(s_{x,w_j}) / \max_{i=1,\ldots,N_j}(score_k(s_{i,w_j}))$$

The weighted votes synset $s_{i,wj}$ receives over all measures is the sum of its weight under individual measure. In voting scheme 2, the synset from the candidate list which gets the highest weighted votes becomes the winner.

$$weightedVote(s_{i,w_j}) = \sum_{k=1}^{K} weight_k(s_{i,w_j})$$

$$s_{w_j} = s_{x,w_j}, if$$

$$weightedVote(s_{x,w_j}) = \max_{s_{i,w_j} \in candidates(s_{w_j})} (weightedVote(s_{i,w_j}))$$

Weighted Votes among All Synsets

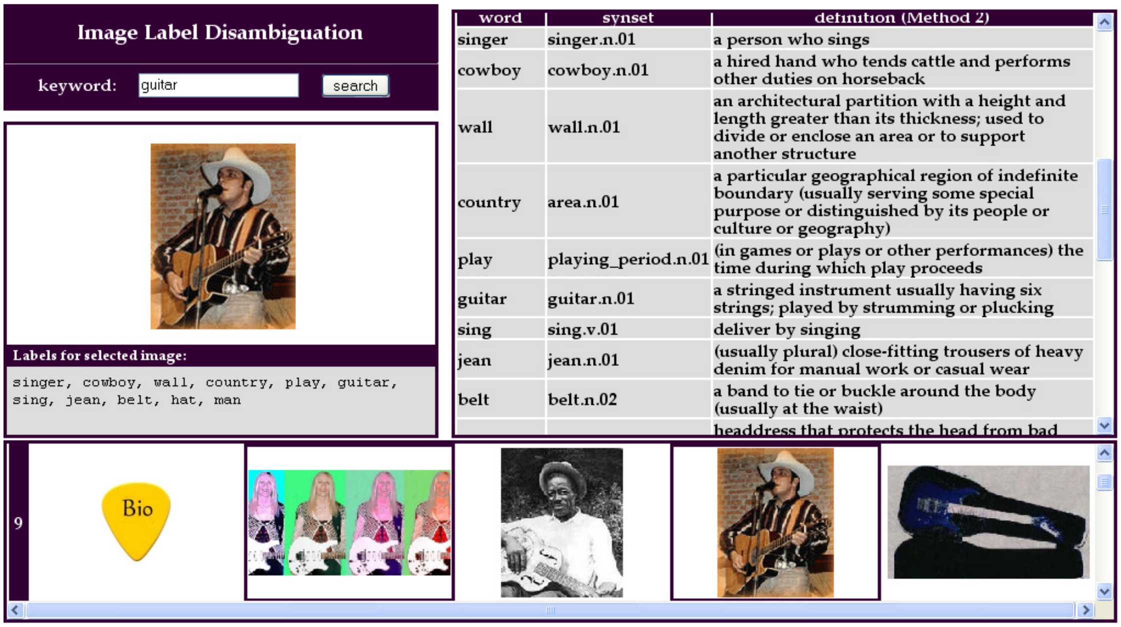
Voting scheme 3 differs from 2 in that the synset from all synsets for word w_j which gets the highest weighted votes is the proposed synset for w_j.

$$s_{w_j} = s_{x,w_j}, if$$

$$weightedVote(s_{x,w_j}) = \max_{i=1,\ldots,N_j} (weightedVote(s_{i,w_j}))$$

Given that the accuracy of WSD cannot reach 100%, our group manually examined the concepts automatically assigned to the Peekaboom images by the voting algorithm. Evaluation results showed that the voting algorithm achieved over 80% accuracy rate, which was significantly higher than any of the existing single-measure WSD algorithm. Figure 4-9 shows the web interface for viewing the label disambiguation results for the Peekaboom images. The nine out of eleven labels were correctly assigned to the concepts denoted by the image content (the two wrong assignments are “country” and “play”). In future research, the manually corrected image label WSD results can serve as training sets for semi-supervised or supervised algorithms. The expansion of the evocation dataset can further help to achieve better WSD performance as well.

As shown in Figure 4-9, more than one picture has a label "guitar." Their contents as well as image qualities are quite different from one another. This is a potential disadvantage for web images to be used as representations in communication support, compared to carefully crafted icons. In the next section, I described two studies that evaluated the efficacy of web images for evoking commonly used concepts in comparison to stylized icons.



Labels for selected image:

singer, cowboy, wall, country, play, guitar, sing, jean, belt, hat, man

word	synset	definition (Method 2)
singer	singer.n.01	a person who sings
cowboy	cowboy.n.01	a hired hand who tends cattle and performs other duties on horseback
wall	wall.n.01	an architectural partition with a height and length greater than its thickness; used to divide or enclose an area or to support another structure
country	area.n.01	a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography)
play	playing_period.n.01	(in games or plays or other performances) the time during which play proceeds
guitar	guitar.n.01	a stringed instrument usually having six strings; played by strumming or plucking
sing	sing.v.01	deliver by singing
jean	jean.n.01	(usually plural) close-fitting trousers of heavy denim for manual work or casual wear
belt	belt.n.02	a band to tie or buckle around the body (usually at the waist)
		headdress that protects the head from bad

Figure 4-9. Example ("guitar") of image label disambiguation results.

4.3 Images vs. Icons: Evocativeness of Static Visual Representations

In *Section 4.1 Static Visual Representation Perception and Usage*, I introduced the differences between the visual cues provided by icons (graphic symbols) and those by images (realistic photos). Both anecdotal reports from speech language therapists and systematic investigation suggest that, unlike written words and abstract symbols, no training is required for understanding images; in fact, experiments have found that infants are able to understand images without training (Hochberg1962).

Ryan et al. (Ryan1956) 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. Their 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 and people with language barriers, a set of studies illustrating that the web images are, on the whole, as effective as icons for communication need to be conducted. Thus, in my thesis research, I have directly compared the effectiveness of web images and icons for communication with both healthy aging population and aphasic individuals.

4.3.1 Research Questions and Challenges

Unlike earlier studies, the vocabulary I examined with images and icons was not restricted to concrete nouns, but extended to illustrate different parts of speech, including concrete and abstract nouns, verbs, and adjectives. Moreover, the source of image representations is relatively random. It is compiled from the Internet and not painstakingly annotated. The goal is to demonstrate that this cheap and plentiful source of images is “good enough” to use in communication systems developed for people with language deficit. The research questions that I would like to address in this research are as follows:

(1) Are web images are effective as stylized icons in conveying frequently used concepts for general population and individuals with language disabilities? The answer to this question will justify if web images can be incorporated into AAC vocabularies.

(2) What kinds of concepts can be easily illustrated by static pictures? Previous research mostly emphasized visual representations on concrete words (especially nouns). I would like to explore if static pictures have the same efficacy in conveying other parts of speech that are essential for communication, such as verbs, adjectives and adverbs.

(3) What affects people's comprehension of static visual representations? If people fail to recognize a given visual representation, it is possible that the selected picture does not provide sufficient unambiguous information. Exploring the factors that cause miscomprehension can help generate guidelines for creation of static visual AAC vocabularies.

To truly understand the communicative power of images and icons, they need to be evaluated with people with aphasia, one of the ultimate user populations. However, working with a group of people with communication disorders is very challenging. First, it is hard to articulate information to people have language comprehension deficit, and also hard to understand what people intend to say if they cannot produce fluent and meaningful speech. Second, at the early stage of a research, the preparation of the data and design of the experiment could have many pitfalls. Presenting such research to people with language disabilities for feedback could be problematic. These people have already struggled with everyday life, and it is big burden for them trying to overcome any flaws in the design of the study. Fourth, without valid gold standard, it is hard for us to tell if the failure of the visual representations is caused by poor design or by the cognitive disability of the participants. Fifth, people with language disabilities tend to withdraw from social life, and are difficult to recruit as participants for a study.

To overcome these difficulties, I first evaluated the effectiveness of the visual representations with healthy elderly individuals. Of cause, stroke and other brain damage may result in cognition impairment, which can affect how people with language disorder perceive information. However, from a broader perspective, most of aphasic individuals fall in the elderly age span, and it is likely that they share similar life experience and interests as well as are affected by similar age-related degeneration as ordinary senior citizens. The results with healthy aging population can serve as a comparison standard for the analysis of the performance of aphasic individuals.

4.3.2 Static Picture Study 1 (Cued Comprehension) with Healthy Aging Population: Study Descriptions

4.3.2.1 Experimental Design and Hypotheses

In the first study, the participants (senior citizens) were asked to interpret visually illustrated words and phrases in which either images or icons were used. The goal was to investigate how well people understand concepts presented in commonly used sentences context. The study was a 2 by 4 factorial design. The independent variables are visual stimuli (web images vs. stylized icons) and parts of speech (concrete nouns, abstract nouns, verbs, and adjectives/adverbs). The dependent variables are response accuracy at word level and at sentence level. The study used a between-subject design, because senior participants tire readily requiring their testing period to be short. Their attendance at the recruitment site was also sporadic, making it difficult to schedule them for another session.

The hypotheses for the study were:

H1. Images in general are better than icons as visual representations for senior citizens.

H2. Concrete nouns are significantly easier to illustrate and interpret than other parts of speech.

H3. Icons are better representations for adjectives and adverbs than images, because of the use of visual indicators.

4.3.2.2 Data Preparation

The procedure for eliciting the meanings of frequently used sentences and phrases worked as follows. A set of sentences relevant to the participant group were collected, and they were transformed into a mixture of either words and images or words and icons. The participants were asked to reformulate the original intended meaning of the given illustrated sentences.

Twenty-five sentences were collected for the experiment: 14 from blogs of senior citizens featured on the Ageless Project and 11 from daily phrases suggested by speech-language pathologists (SLPs). The Ageless Project (ageless2006), an online blog forum for elderly people, shows the common topics among senior citizens. Speech-language pathologists, who are experts working with people with aphasia for language rehabilitation, understand the types of phrases that are the most useful in the day-to-day life of aphasic individuals', our ultimate target population. Table 4-1 and 4-2 show the sentences used in the study from blogs and the SLPs, respectively. The underlined words in these sentences were replaced by pictures.

1	I've wished and dreamt of having a talent to perform.
2	We started to have car trouble, but we weren't too far from a service station.
3	My eye drops fell in the toilet.
4	I'm buying a vacation home.
5	My ice cream is made from milk, strawberry sauce, and white chocolate.
6	The therapist wants to schedule me for a day a week.
7	While putting on pants, I was unable to stand on my right leg.
8	I opened a bank account, and then phoned the Social Security office.
9	I had to shovel snow from my car, and I drove.
10	I'm not happy with the pain medication at night.
11	I was surprised that all seats in the House are up for election.
12	I am taking a class and need to bring pictures to the first class.
13	My family didn't have money for school.
14	I don't have a fishing license because the government can't tell me I can't fish.

Table 4-1. Fourteen sentences from the Ageless Project.

15	Can you help me?
16	Can you write that for me?
17	Where is the bathroom?
18	Where is the bus stop?
19	I am hungry.
20	Where is the hospital?
21	I went to a football game.
22	I fell down and broke my leg.
23	I lost my little dog. It was hiding behind the flower pot.
24	We celebrated my wife's birthday in the backyard.
25	I made an appointment with my dentist this week.

Table 4-2. Eleven sentences suggested by the Speech-Language Pathologists.

The sentences from elderly people's blogs often had to be paraphrased. While the paraphrased ones are by design specifically transactional, i.e., carrying message content information, the originals often employ idiosyncratic, interactional content that promotes a closer connection with the reader. For example, the original sentence "Speaking of things to thank aging for, this morning as I was dressing, I found I was unable to stand on the right leg, while putting my left into the pants leg." was simplified as "While putting on pants, I was unable to stand on my right leg.'" Although some information is lost, this study focuses on the information-carrying ability of pictures rather than the emotional component.

Paraphrasing also helps to increase the ratio of pictures to words in the phrases because I replaced only words (nouns, verbs, and adjectives) in the phrases that can be illustrated by icons and images. This preference thus biased the experiment toward more of the common terms that were in both resources. The variety of sentences allowed us to represent both short and long sentences. The shortest sentence was three words and the longest was 15. Likewise, there was a range of substitution; the ratio of the number of pictures to the total length of the sentence went from 0.17 to 0.57, thus ensuring that even in the most difficult sentences there remained a syntactic context for the subjects to use.

For each of the selected words, the appropriate image was chosen using the following process. Our group grabbed all Peekaboom images that were assigned to concepts (see *Section 4.2*) appearing the tested phrases. This step ruled out images where the word represented a different sense than that conveyed by the image. Images of low quality were rejected as well. At the end, three judges voted separately for the best image. Online image search results were added if not enough images were retrieved from the Peekaboom 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. For a more concrete example, the illustrated phrase “I need to bring pictures to the first class.” is shown in Figure 4-10. It includes image and icon representations for a concrete noun “picture,” an abstract noun “class,” a verb “bring” and an adjective “first.” As one can see, the size and quality of the web images varied a lot, in contrast of icons of a consistent style.

This study was co-conducted with my colleagues Jordan Boyd-Graber Ying and Sonya Nikolova.

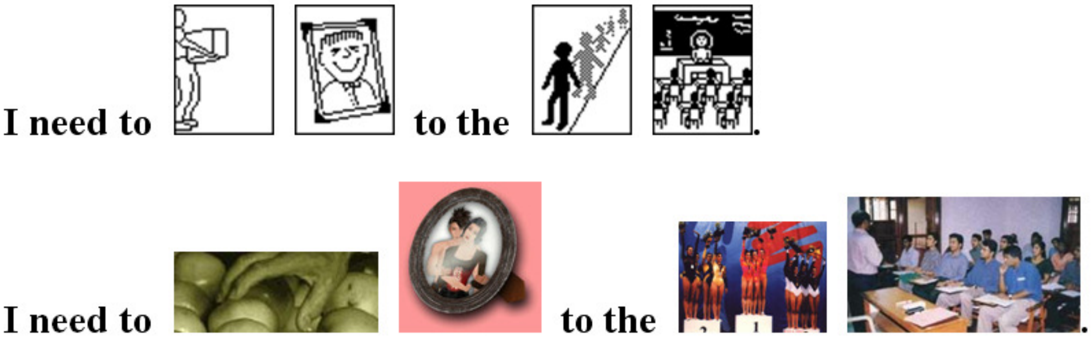
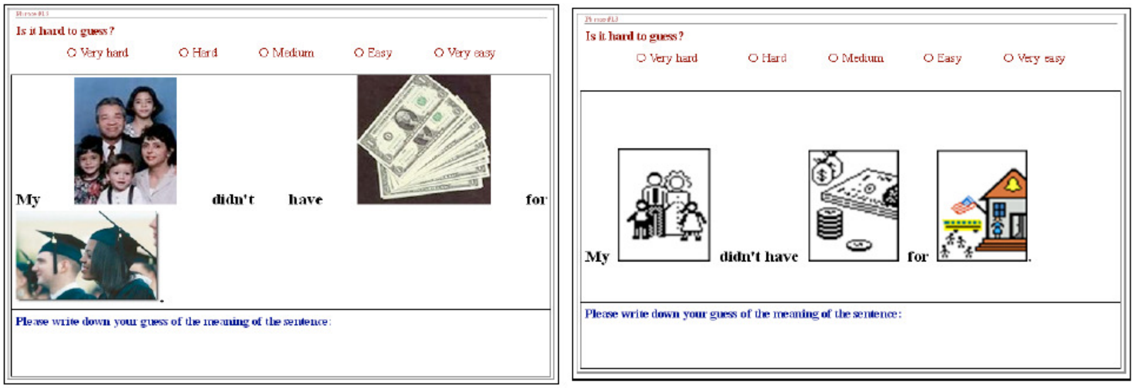


Figure 4-10. Content words replaced with icons (top) and images (bottom).

4.3.2.3 Study Procedure and Participants

The phrases were presented to individuals in a booklet form with one phrase to each sheet. We asked the participants to write down their guess of the whole sentence 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, median, easy, and very easy). Figure 4-11 shows examples of image and icon testing sheets (“My family didn't have money for school.”), respectively. At the top of each sheet is a place for the user to record the perceived difficulty, the middle is the phrase with words substituted, and the bottom has a field for the user to write the complete sentence.



Phrase #13

Is it hard to guess?

○ Very hard ○ Hard ○ Medium ○ Easy ○ Very easy

My didn't have for .

Please write down your guess of the meaning of the sentence:

Phrase #13

Is it hard to guess?

○ Very hard ○ Hard ○ Medium ○ Easy ○ Very easy

My didn't have for .

Please write down your guess of the meaning of the sentence:

Figure 4-11 Example of testing sheet.

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-based 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 experiments, every participant picked a number (1-24) from a hat, and the booklet with corresponding index was given to the participant. They were then read the following instructions:

> You will be given 25 sentences, but some words are missing. These words have been replaced by pictures chosen to represent the word that has been replaced. At the bottom of each page is a space for you to write down what you think the complete sentence is. After you have written down your answer, please rate how difficult you thought it was to reconstruct the sentence by circling one of the five choices at the top of the page. Please work as quickly as you can, but please do not rush so much as to sacrifice accuracy.

When finished, the investigator sealed the booklet so that only the answer part of the booklets could be viewed. The answers were transcribed into text files.

Twenty-four participants were recruited though posting advertisement at the Princeton Senior Resource 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. Participants all have high school and above education. Magnify glasses were provided in case of age-related vision degeneration.

4.3.3 Static Picture Study 1 (Cued Comprehension) with Healthy Aging Population: Results and Analysis

4.3.3.1 Scoring Scheme

In order to evaluate how well pictures conveyed the meaning of an intended sentence, I need to quantify the degree to which two pieces of text differ. This is also a problem in evaluating outputs in machine translation, so the General Text Matcher (GTM, 2006) was used initially, 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. Table 4-3 shows the top five phrases with the greatest disparity between the performance in conveying sentence meaning with the icons and images based on the median test; the stimulus with the higher score is in bold, and the scores for sentences as a whole is presented for reference. As shown, only one phrase (#18), “Where is the bus stop?” had significantly different scores.

Phrase#	Image	Icon	F(1, 22)	p-value	η^2 effect size
18	0.704	**0.983**	8.358	**0.008**	0.275
6	**0.684**	0.551	2.656	0.117	0.108
15	0.687	**0.803**	2.491	0.129	0.102
11	0.531	**0.661**	1.460	0.240	0.062
4	**0.507**	0.407	1.368	0.255	0.059

Table 4-3. Sentences with the greatest disparity between responses given icons and images by ANOVA on the GTM scores.

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, the investigation was refined using a more precise error metric based on human judgment.

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 4-4. 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.

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”

Table 4-4. Human Scoring Scheme (example: “dog”).

Table 4-5 shows that the human-based scoring system, when rating the sentence as the average of the words in the sentence, gives results comparable to the machine-based method. The human-based method, however gives us the ability look at word-level comprehension effects.

Phrase#	**Image**	**Icon**	**F(1, 22)**	**p-value**	**η^2 effect size**
15	3.833	**4.958**	6.578	**0.018**	0.230
18	4.208	**5.000**	4.538	**0.045**	0.171
21	**4.958**	4.333	3.052	0.095	0.122
13	**4.583**	4.000	2.042	0.166	0.085
12	2.333	**2.875**	1.202	0.285	0.052

Table 4-5. Sentences with the greatest disparity between responses given icons and images by ANOVA on the human scores.

4.3.3.2 Word-level Effects: Visual Stimuli and Parts of Speech

There were 33 concrete nouns, 17 abstract nouns, 10 adjectives and 24 verbs tested in our study. ANOVA was applied to the human scores for each intended word, comparing the accuracy difference between the icon and image mode. Test of homogeneity of variances showed that the responses came from the same normal distribution.

Figure 4-12 shows that there is no significant difference between the performance of icons and images at word level in general (F(1, 22) = 0.050, p = 0.825, η^2 = 0.002). When looking into each part of speech, still, the effectiveness of icons and images in illustrating concrete nouns (Figure 4, F(1, 22) = 0.077, p = 0.785, η^2 = 0.003), abstract nouns (Figure 5, F(1, 22) = 0.068, p = 0.796, η^2 = 0.003), verbs (Figure 6, F(1, 22) = 0.732, p = 0.401, η^2 = 0.032), and adjectives (Figure 7, F(1, 22) = 1.115, p = 0.302, η^2 = 0.048) is not significantly different.

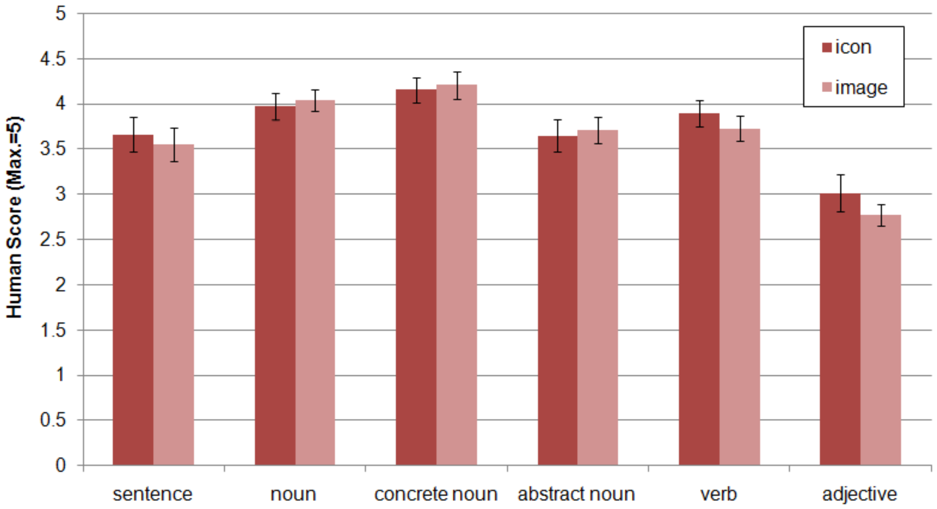


Figure 4-12. Comparison of scores between icons and images across different parts of speech.

However, both images and icons performed equally poorly in capturing the essence of adjectives. Adjectives tested in the study, such as “white”, “first”, and “little”, were rarely correctly interpreted and seemed to be the most likely cause for confusion in interpreting images, as participants tried to cast pictures representing 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.

To identify the concepts where icons and images had disparate effectiveness in communicating the underlying meaning, all of the scores for each word across all subjects and performed an ANOVA between the image and icon groups were examined. Other than the ones listed in Table 4-6, there is no significant difference in the median scores for all target words between modes.

Word	Image	Icon	F(1, 22)	p-value	η^2 effect size
help (v.)	3.625	**5.000**	7.548	0.012	0.255
house (n.)	**4.750**	3.380	5.829	0.025	0.209
vacation (n.)	**4.708**	3.750	4.666	0.042	0.175
first (adj.)	1.250	**2.125**	3.769	0.065	0.146
stop (n.)	4.880	**5.000**	3.667	0.069	0.143

Table 4-6. Words with the Greatest Disparity in Guesses between Icon and Image (ANOVA).

Given that icons are usually pictures of things, it was often 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 too was reflected in how subjects responded to images. It might imply that part of speech as one of the lexical features, could affect the performance of different visual representations. Thus, further analysis on word features was carried out.

4.3.3.3 Word-level Effects: Other Linguistic Properties

In order to further investigate what kinds of concepts can be effectively conveyed by static pictures, I conducted post-hoc tests on various linguistic properties. I selected concreteness, imageability, and familiarity (based on the MRC Psycholinguistic Database) since they are directly or indirectly related to visualizability of the concepts. The result of ANOVA (Table 4-7) shows that no word features other than part of speech have significant effect on the interpretations, regardless of the mode of visual representations. While overall they did very well in identifying nouns, subjects had more difficulty with adjectives and verbs. However, a linear trend was found in both concreteness and imageability; the more concrete/imageable the concept is, the more accurate people can interpret it from visual content (Figure 4-13). For concepts that are more abstract or less imageable, icons performed slightly better than images; on the contrary, images worked slightly better than icons with more concrete or more imageable concepts.

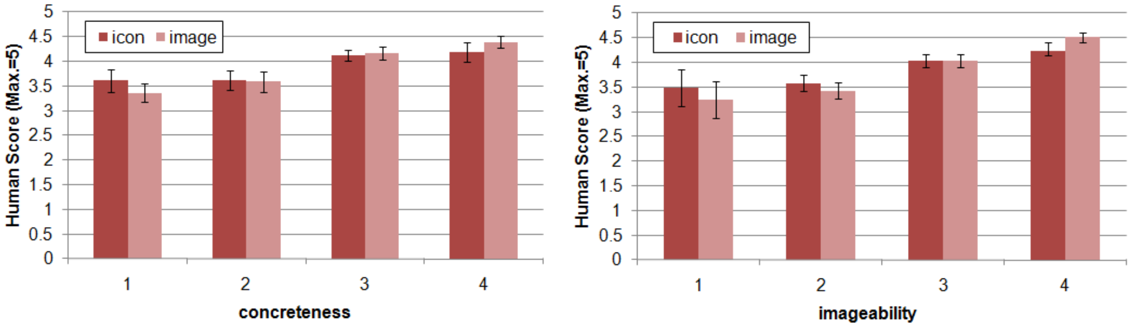


Figure 4-13. Change of human score with concreteness (left) and imageability (right).

stimuli	df	F	p-value	η^2 effect size
parts of speech (N=41)				
overall	3	4.027	0.013	0.228
icon	3	3.602	0.021	0.209
image	3	3.597	0.021	0.208
concreteness (N=41)				
overall	3	1.018	0.395	0.069
icon	3	1.250	0.304	0.084
image	3	0.764	0.521	0.053
imageability (N=41)				
overall	3	1.191	0.325	0.080
icon	3	1.083	0.367	0.073
image	3	1.136	0.346	0.077
familiarity (N=42)				
overall	2	0.695	0.505	0.033
icon	2	0.763	0.473	0.036
image	2	0.486	0.619	0.023

Table 4-7. Influence of various linguistic properties given icons and images.

4.3.3.4 Sentence-level Effects: Visual Stimuli and Parts of Speech

We also looked into how people interpreted not only the meaning of individual concepts, but also the key idea of the whole sentences. The ANOVA results showed that accuracy in sentence interpretation (Figure 4-12) had no significant difference between icons and images ($F(1, 22) = 0.301$, $p = 0.589$, $\eta^2 = 0.013$). Table 4-5 shows that Phrase 15 ("Can you <u>help</u> me?") and Phrase 18 ("Where is the <u>bus</u> <u>stop</u>?") have significant disparity between responses given icons and images based on human scoring. Although both "help" and "stop" are among the words which have great disparity between image and icon mode, it is likely that sentence features of those phrases consisted of words and pictures, such as length, absolute number of pictures, and the picture-word ratio, have something to do with the differences. We are interested in investigating how well those features affect people's responses.

The four phrases with both average icon and image scores under 3 all have words with individual average score below three in them: Phrase 2 ("trouble," "start"), Phrase 7 ("right"), Phrase 23 ("lose," "little"), and Phrase 12 ("first"). It implies that even with the context, if some key words were misinterpreted, the meaning of the sentence would not be conveyed correctly.

4.3.3.5 Sentence-level Effects: Other Sentence Features and Context

In order to have a better understanding of the possible influence that sentence features, such as the length of the sentence, number of pictures in the sentence, and picture-to-word ratio, may have on people's interpretation performance, post-hoc tests were conducted. The ANOVA results showed that none of the sentence features had significant impact. However, for icons, the length of the sentences, and for images all the features had a great effect size (greater than 0.140), which means the impact might be significant if more participants were involved (Table 4-8). Figure 4-14 shows that as the length of a sentence increase to more than 11 words, the accuracy drops greatly. It is also shown that though not significantly different, as the number of pictures (Figure 4-15) or the picture-to-word ratio (Figure 4-16) goes up, people's interpretation accuracy decreases.

There are six words appearing twice in the test, with some of repetition in the same phrase and some not. The words "car," "leg," "week," and "fall" appear in different sentences, while "class" and "fish" appear in the same sentence. Though there is no significant difference (Figure 4-17) between the two responses for each word but "car", there is some slight variance, which means context may affect the interpretation to some extent.

stimuli	**df**	**F**	**p-value**	**η^2 effect size**
length of the sentence (N=22)				
overall	3	0.886	0.484	0.228
icon	3	0.797	0.526	0.210
image	3	1.468	0.288	0.329
number of embedded pictures (N=22)				
overall	3	0.534	0.671	0.151
icon	3	0.309	0.818	0.093
image	3	0.929	0.466	0.236
picture-to-word ratio (N=22)				
overall	3	0.920	0.433	0.170
icon	3	0.218	0.808	0.046
image	3	2.530	0.134	0.360

Table 4-8. Influence of various sentence features given icons and images.

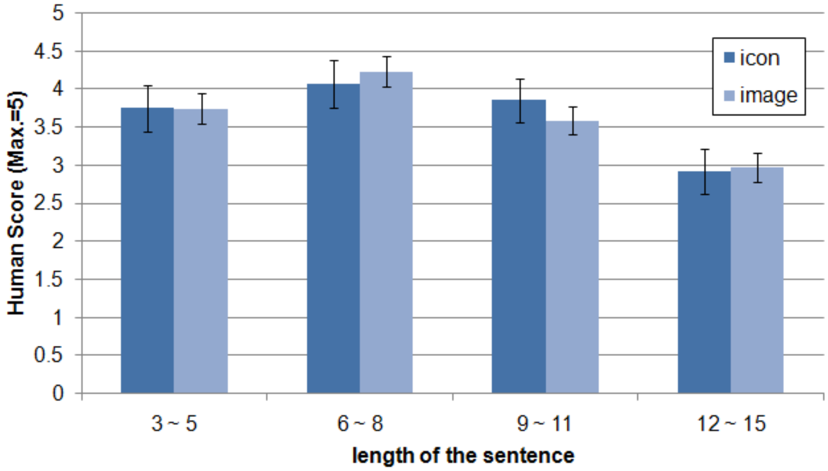


Figure 4-14. Influence of the length of the tested sentences.

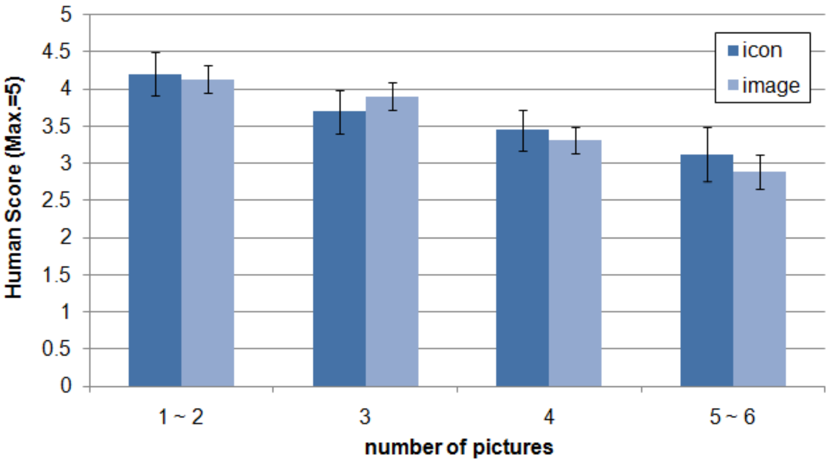


Figure 4-15. Influence of the number of pictures in the sentences.

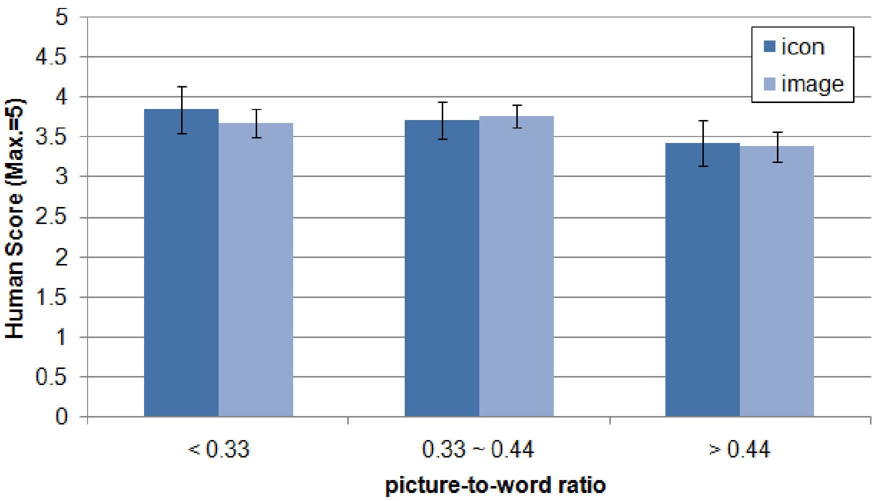


Figure 4-16. Influence of picture-to-word ratio in the tested sentences.

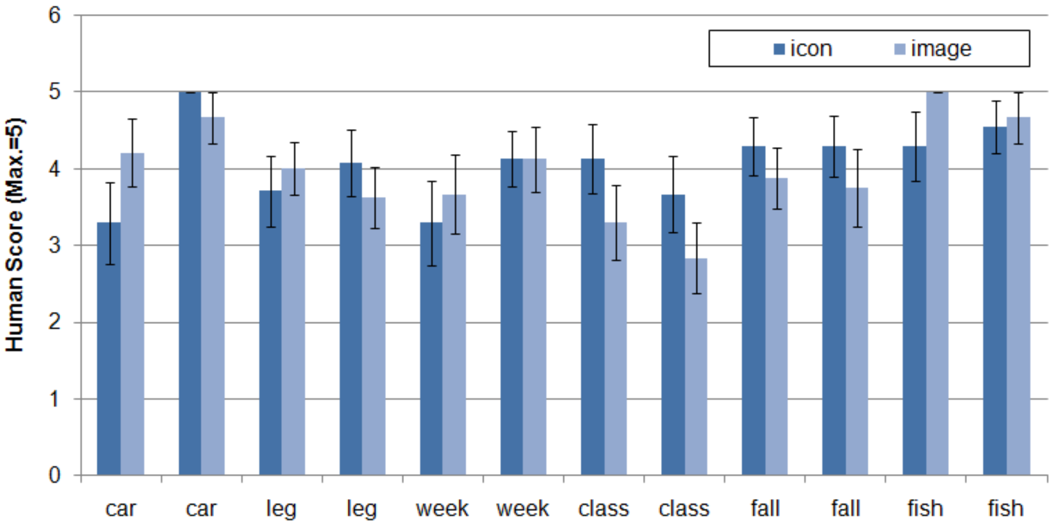


Figure 4-17. Influence of context on words appearing in multiple sentences.

4.3.3.6 Aggregate Effects

When I computed aggregate scores for each individual on an entire booklet, I did see some slight differences between those who saw images and those who saw icons. Figure 12 is a histogram showing the total score (taken as the sum of the accuracies of each individual word and sentence) for both of the stimuli. While the mean of the scores for the icon subjects (412) was higher than the scores of the image subjects (392), the variance for each group (44 and 53, respectively) was great enough that the null hypothesis that the means are equal could not be rejected under a T-test.

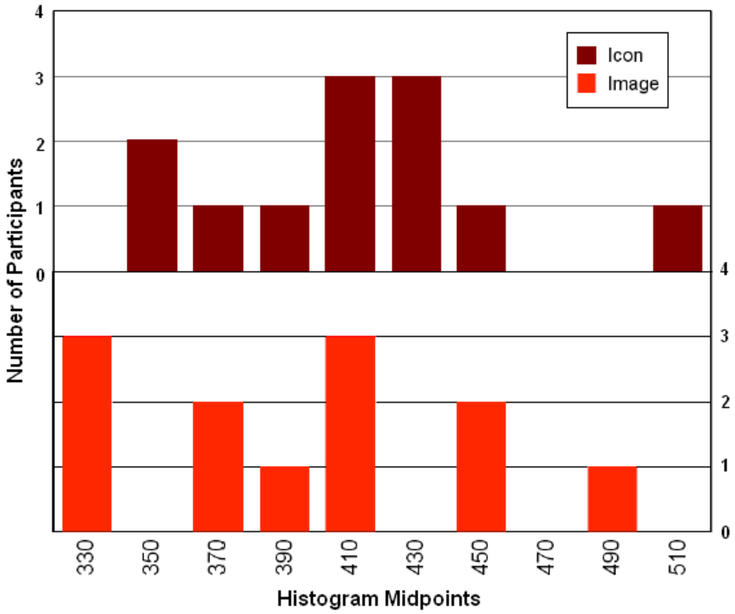


Figure 4-18. Histogram of the total scores per participant.

4.3.3.7 Age, Gender, and Self-reported Interpretation Difficulty

The correlations between age and performance and between gender and performance were both negligible. There was a slight correlation between age and the overall time spent on the test, which could simply be a result of slower handwriting or poorer vision. While those with the highest scores took the longest time, the correlation between time spent and the accuracy was weak.

The self-rated difficulty of the two stimuli were approximately the same, although the images were perceived as being slightly less difficult (a mean rating of 3.31 versus 3.37), but this difference is not statistically significant. Because many subjects did not rate the difficulty and left the difficulty field blank for different reasons (some thought that really easy sentences were not worth rating, and others would leave whole sheets blank when stumped), any conclusions based on the difficulty data would be spurious.

4.3.3.8 Conclusions

In summary, for senior citizens, there was no significant difference between the effectiveness of images and that of icons as concept carriers (hypothesis H1 rejected). Nouns, especially concrete nouns are significantly easier to convey via a static picture than other parts of speech (H2 confirmed). However, the performance difference between images and icons within each part of speech was not significantly different (H3 rejected). Linguistic properties such as concreteness, imageability, and familiarity as well as sentence features did not show significant impact on the evocativeness of the visual stimuli.

4.3.4 Static Picture Study 2 (Noun Naming) with People with Aphasia: Study Descriptions

The study that compared web images and stylized icons as visual representations with healthy elderly adults showed that web images, despite the variation in quality, were as effective as deliberately designed icons. Since one of my ultimate goals is to explore economic and effective visual representations and apply them to assistive technologies for communication for aphasic individuals, a second study was designed and carried out with participants from the target population in order to assess these two paradigms for conveying semantic information, nouns in particular, visually at word level. The parts of speech were narrowed down to nouns only because other concepts such as verbs and adjectives are often hard to visualize and can be confusing given only visual representations even for an unimpaired population.

4.3.4.1 Experimental Design and Hypotheses

In the second study, both icons and images were presented to individuals affected by aphasia. The participants were asked to choose, from a list of five words, the word that best represents the meaning conveyed by the picture. Because the extreme variability of impairment would prevent us from making comparisons across individuals, we were forced to show both icons and images to each participant. In addition, in our previous study with senior citizens, a population more susceptible to acquire aphasia than other age groups, it was shown that images and icons have reasonably good performance only in illustrating nouns among all parts of speech. We decided to focus on noun concepts which are also the most significant component in communication. The study was a 2 by 2 factorial design. The independent variables are visual stimuli (web images vs. stylized icons) and parts of speech (concrete nouns and abstract nouns). The dependent variables are response accuracy, response diversity, and response time.

The hypotheses for the noun naming study with aphasic individuals were:

H1. Images are as effective (in terms of response accuracy) as icons on illustrating nouns for people with aphasia.

H2. Images are as effective (in terms of response diversity) as icons on illustrating nouns for people with aphasia.

H3. Images and icons take similar amount of time for aphasic individuals to interpret.

H4. Concrete nouns are significantly easier to illustrate and interpret than abstract nouns.

H5. It takes significantly shorter time to interpret concrete nouns than to interpret abstract nouns.

4.3.4.2 Data Preparation

Twenty-five nouns (Table 4-9), of which nine are abstract (underlined in the table) and 16 are concrete, were selected from sentences tested in the first study with senior citizens. Nouns represented by low quality images were removed.

For each word, an icon or an image representing the chosen noun was displayed in addition to five options, including the target and four other alternatives. The icons came from Lingraphica, and the images came from the Peekaboom Image Label Dataset, which were inherited from the previous study with elderly people. To avoid bias on the content of the pictures, we did not include any noun that is illustrated by completely different icon and image contents.

The alternatives of each noun were chosen using the following scheme.

(1) WordNet: sister terms (other direct hyponyms of its direct hypernym), part meronym, and uncle terms (sister terms of its direct hypernym). The definitions of the terminologies see *Section 2.1*.

(2) Peekaboom labels: only when not enough alternatives are found in (1), and the icon and image for the target look similar.

(3) Guesses from the previous study (the icon/image study with senior citizens): when not enough alternatives are found in (1) and (2).

The chosen alternatives can replace the target noun in a sentence without violating selection restrictions, which means that the alternatives can be used in the same or similar context. The alternatives were selected manually from the long list of candidates and reviewed by speech-language pathologists to ensure that the words are not too difficult for aphasic individuals.

4.3.4.3 Study Procedure, Interface, and Participants

The previous study was paper-based and asked elderly participants to write down sentences in which words had been replaced by pictures. However, in this study, we could not rely on the ability of aphasic participants to understand written words that form a word's context nor could we rely on users having the verbal and motor skills necessary to write complete sentences. Thus, we used multiple-choice answers to restrict the answer space, make input easier, and to provide an alternative means of limiting the semantic answer space now that context is unavailable.

The study was conducted on a web-based interface (Figure 4-19). The layout favors the left-hand side of the screen, which 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 speech files, to offer consistent auditory clues for subjects who are unable to read individual words. It was ensured that each page does not exceed a single screen, thus eliminating the need to scroll. Users were allowed to signal their decisions by either directly manipulating the mouse on the screen, or saying the answers out loud.

Intended Noun	Options (alphabetic order)
appointment	appointment; diary; holiday; luncheon; session
birthday	anniversary; birthday; Christmas; holiday; wedding
car	aircraft; car; motorcycle; trolley; vehicle
class	assembly; camp; class; crowd; room
dentist	dentist; doctor; nurse; pharmacist; veterinarian
dog	bear; cat; dog; fox; wolf
eye	ear; eye; lashes; organ; pupil
family	children; family; folks; group; team
flower	flower; fruit; leaf; shrub; tree
government	bureaucracy; government; law; party; white house
house	building; dormitory; hotel; house; lodge
ice cream	bowl; dumpling; ice; ice cream; sundae
leg	arm; body; fin; foot; leg
medicine	elixir; food; injection; medicine; treatment
money	check; money; salary; stamp; wealth
night	day; hours; morning; night; season
pants	lap; pants; shirt; skirt; suit
picture	art; drawing; frame; map; picture
pot	basket; bowl; can; pan; pot
seat	bed; seat; sofa; stool; table
strawberry	apple; banana; grape; peach; strawberry
toilet	bathroom; shower; sink; toilet; urinal
trouble	boredom; excitement; fun; luck; trouble
vacation	journey; pilgrimage; quest; vacation; visit
week	day; fortnight; month; week; year

Table 4-9. Twenty-five intended nouns (underlined ones are abstract) with five options.

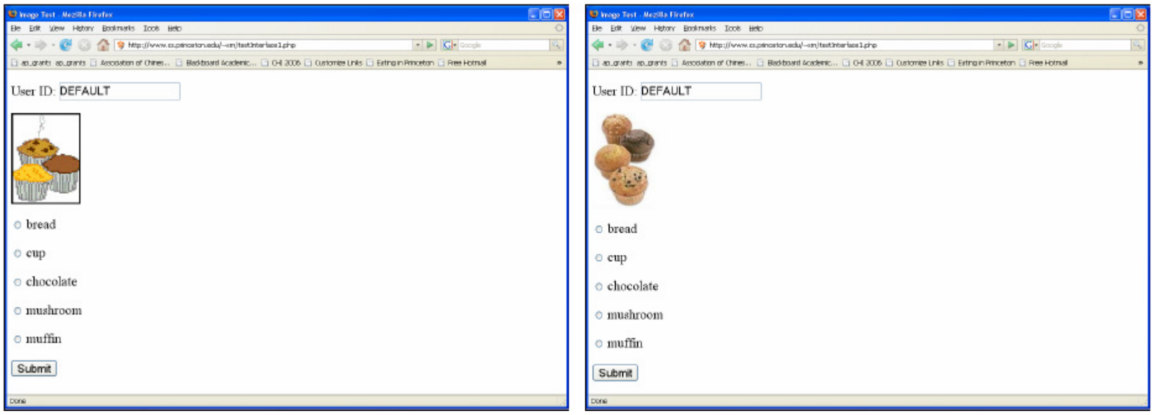


Image Test - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Submit
Done

User ID: DEFAULT

○ bread

○ cup

○ chocolate

○ mushroom

○ muffin

User ID: DEFAULT

○ bread

○ cup

○ chocolate

○ mushroom

○ muffin

Figure 4-19. Web interface for the noun naming study with people with aphasia.

Each participant had to go through all the 25 nouns, half of which were represented by icons and the other half by images. The testing order of nouns, which visual representation to present first, and which nouns went with which visual mode were randomly assigned. Each selection was verbally confirmed before it was submitted and sent to a central server, which also recorded the time between answers.

Fifty aphasic individuals were recruited from a local aphasia center for the study, including 15 females and 35 males. 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.

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 familiarity with computer use before they had aphasia, all were now taking computer lessons through the center. The coordinators of the center provided three broad groupings of impairment based on computer based assessments: 27 of them are high functioning, meaning they have little difficulty in linking pictures to words; 18 are median functioning, meaning they can recognize some of the words, but not the whole sentence; and 5 are low functioning, meaning they have trouble even indentifying single words.

This study was co-conducted with my colleagues Jordan Boyd-Graber Ying and Sonya Nikolova.

4.3.5 Static Picture Study 2 (Noun Naming) with People with Aphasia: Results and Analysis

This section seeks to examine differences between the users' response to pictorial stimuli in our trial. We first explore different methods of gauging the accuracy of responses and finding little difference, and then examine whether 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.

4.3.5.1 Evaluation Metrics

Participants’ performances were measured by response accuracy (accuracy rate and error rate), response diversity (entropy), and response time.

Accuracy rate is the percentage of responses that match the intended answer. **Error rate** equals one minus accuracy rate. However, accuracy 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. A desirable means 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. In the ideal case where all the participants chose the same option, the entropy is 0.82. The higher the entropy is, the less people agreed with each other on the concept represented in the picture.

Response time refers to the interval between the time that a visual stimuli and naming options show up on the interface and the time that a selection is made and submitted. For consistency, the investigators operated the mouse and had the participants indicate their answers by pointing. However, some participants had trouble understanding text and needed to listen to the speech sounds for each option. Their response time was greatly extended.

4.3.5.2 Visual Stimuli and Parts of Speech

Test of homogeneity of variances showed that participants’ performance with images and icons came from the same normal distribution.

ANOVA test on accuracy rate, entropy, and response time shows that in general there is no significant performance difference between the image and icon mode (Table 4-10). However, parts of speech (POS) showed a significant effect (Table 4-10), meaning concrete nouns were significantly easier and faster to perceive than abstract nouns, although the impact was similar for icons and images (Figure 4-20).

factor	F(1, 46)	p-value	η^2 effect size
accuracy rate			
POS	5.188	0.027	0.101
stimuli	0.016	0.900	<0.001
POS × stimuli	0.028	0.868	0.001
entropy			
POS	6.903	0.012	0.130
stimuli	0.231	0.633	0.004
POS × stimuli	0.003	0.959	<0.001
response time			
POS	5.153	0.028	0.098
stimuli	0.816	0.371	0.016
POS × stimuli	0.442	0.509	0.008

Table 4-10. Part of speech (POS), visual stimuli (image vs. icon) and their interaction.

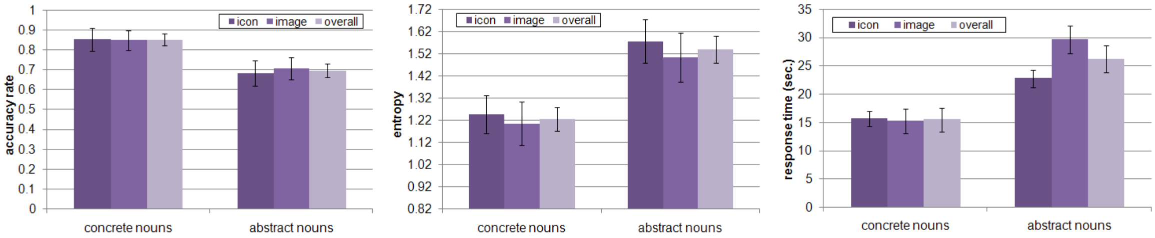


Figure 4-20. Performance of images and icons on abstract and concrete nouns.

Looking at the top five nouns ("appointment," "government," "pot," "trouble," and "week") that people made the most mistakes on (Figure 4-21) and that people mostly disagreed with one another (Figure 4-22) for both pictorial stimuli, four out of five are abstract. These four abstract words (underlined above) were also the ones that took people the longest time on average to perceive (Figure 4-23). Images had lower entropy than icons in 16 out of 25 nouns, meaning that images are more specific in those cases. Entropy had nearly perfect (0.95) correlation with error rate, so subsequent comparisons will be made with the easier to understand metric.

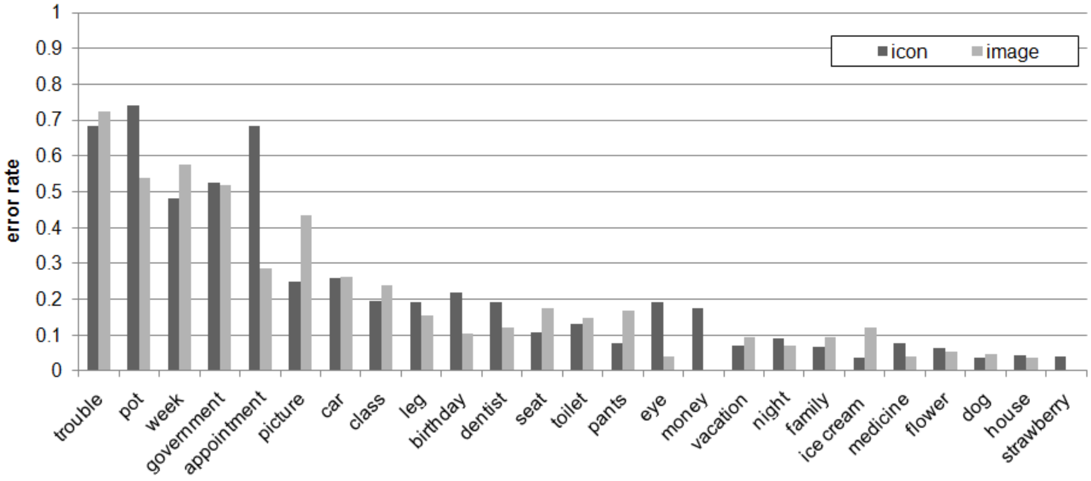


Figure 4-21. Error rate across all tested nouns.

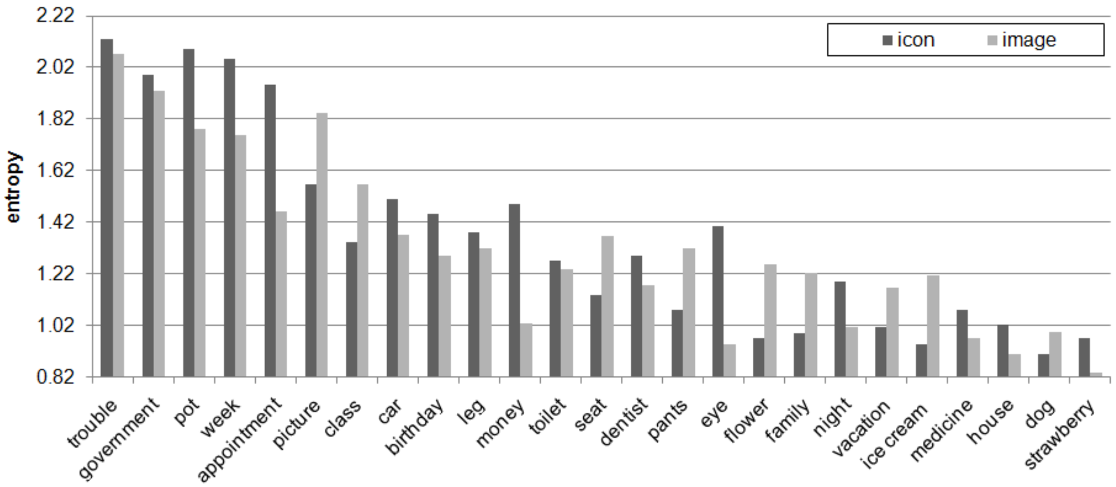


Figure 4-22. Entropy across all tested nouns.

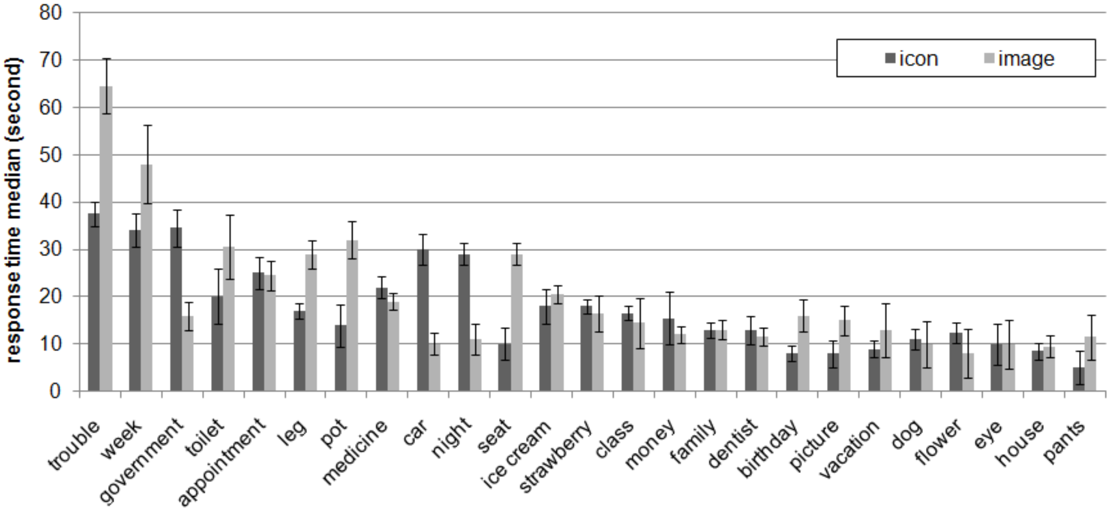
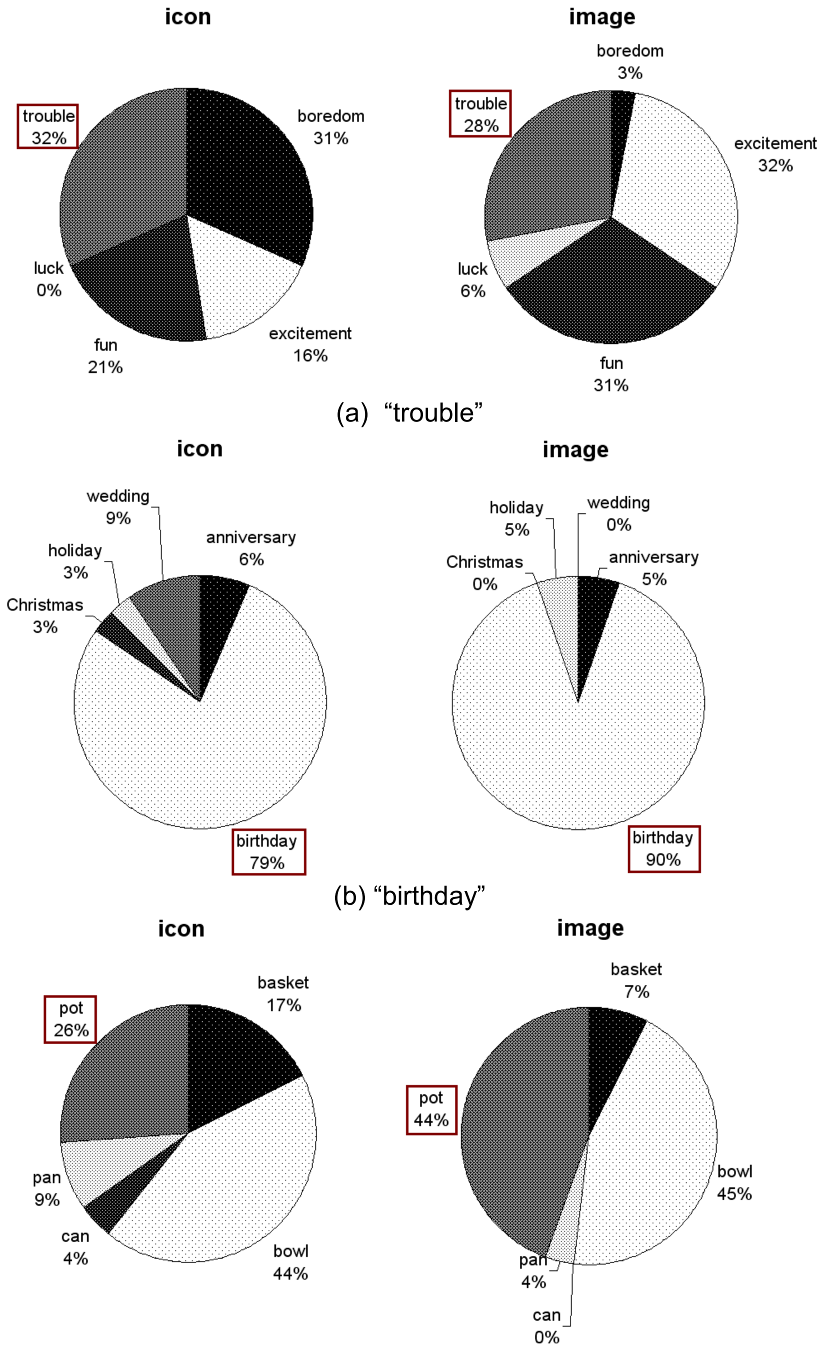


Figure 4-23. Median response time across all tested nouns.

The word “trouble” (Figure 4-24 (a)) is the most ambiguous among all 25 tested nouns for both icons and images. The distribution of responses for “trouble” with icons and that with images have certain similarity. Given the very similar entropy values between images and icons for most of the tested nouns, I further looked into whether the distribution over responses for the two settings is actually different. Table 4-11 shows the top four words of which the distributions are the most different (< 0.005) according to the result of chi-square test. The probability of each option being selected for the top two nouns “birthday” and “pot” are as follows (Figure 4-24 (b)(c)). Although for “birthday”, “pot”, “leg”, and “toilet” there was a statistically significant difference between icons and images, taken in aggregate the data do not support a clear superiority of one or the other.

noun	p-value	error rate (icon)	error rate (image)
birthday	0.0023	0.22	0.11
pot	0.0027	0.74	0.54
leg	0.0038	0.19	0.15
toilet	0.0041	0.13	0.15

Table 4-11. Top four nouns with the biggest response distribution differences (p-value of Chi-square Test) and comparison of their error rate with image and icon representation.



(a) “trouble”

(b) “birthday”

(c) “pot”

Figure 4-24. Examples of response distribution.

4.3.5.3 Other Semantic Properties

For the post-hoc test, once again, the effect of linguistic properties, including concreteness, imageability, and familiarity, was explored. ANOVA results (Table 4-12) showed that the more concrete or more imageable a noun was, the more accurate (Figure 4-25) and faster (Figure 4-26) aphasic people's responses were. Familiarity of the words (based on the MRC Psycholinguistic Database) seemed to have no significant impact on people's interpretation, but in general, the more familiar the words are, the faster people can perceive them from images and icons. No significance was found in the interaction between the linguistic properties and the visual stimuli, which suggests that people's perception of images and icons were affected in a similar way.

linguistic property	df	F	p-value	η^2 effect size
accuracy rate				
concreteness	(3, 46)	13.338	< 0.001	0.225
imageability	(3, 46)	24.765	< 0.001	0.350
familiarity	(2, 46)	0.251	0.619	0.005
entropy				
concreteness	(3, 46)	11.591	0.001	0.201
imageability	(3, 46)	28.233	< 0.001	0.259
familiarity	(2, 46)	0.138	0.712	0.003
response time				
concreteness	(3, 46)	10.095	0.003	0.180
imageability	(3, 46)	10.463	0.002	0.185
familiarity	(2, 46)	0.742	0.394	0.016

Table 4-12. Influence of three linguistic properties (ANOVA).

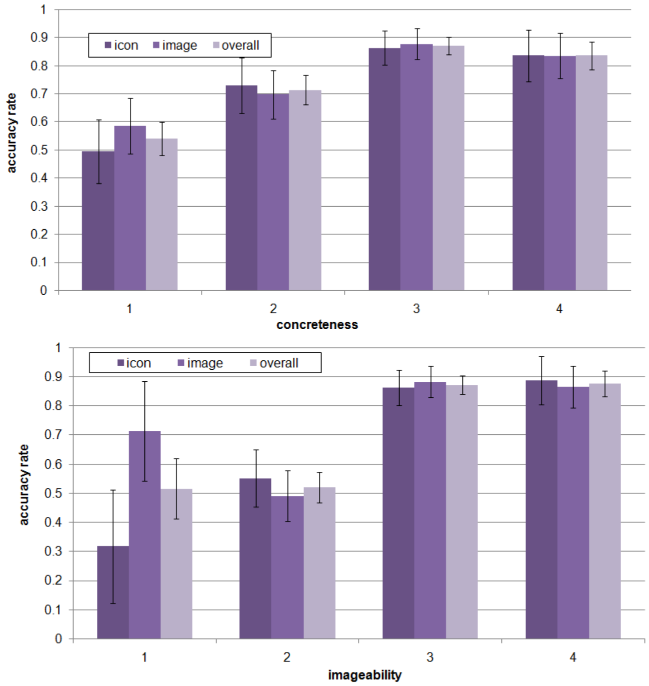


Figure 4-25. Impact of concreteness (top) and imageability (bottom) on response accuracy rate.

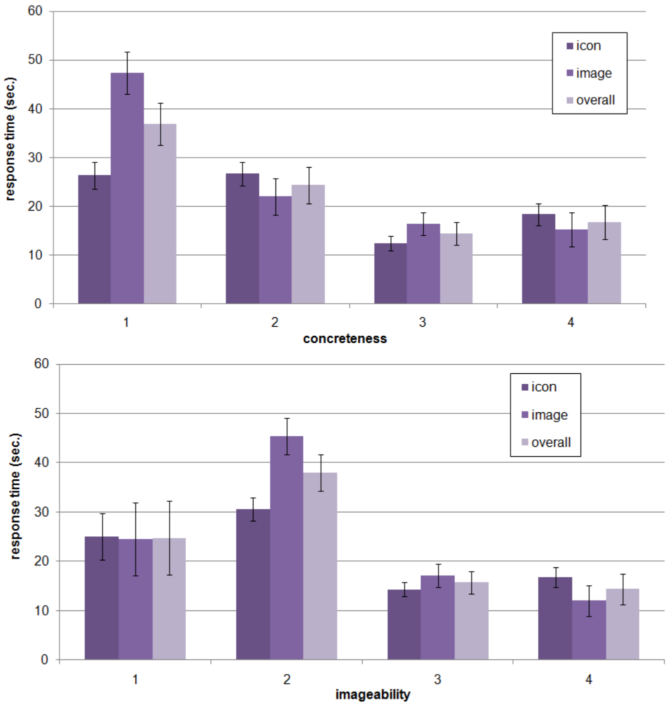


Figure 4-26. Impact of concreteness (top) and imageability (bottom) on response time.

4.3.5.4 Demographic Differences

In general, participants performed similarly with icons and images in terms of both accuracy (F(1, 43) = 1.277, p = 0.265, η^2 = 0.028) and response time (F(1, 44) = 0.266, p = 0.609, η^2 = 0.006), though when looking into demographic variance, there are some differences (evaluated through post-hoc tests).

Females in general are slightly but not significantly better (Table 4-13) at perceiving pictures than males, in terms of both time (F(1, 44) = 2.621, p = 0.113, η^2 = 0.056) and accuracy (F(1, 44) = 1.088, p = 0.303, η^2 = 0.024). However, since only 15 out of 50 participants were female, the conclusion is not solid.

stimuli	accuracy rate			time (second)		
	average	female	male	average	female	male
icon	0.79	0.81	0.78	21.50	17.79	22.73
image	0.79	0.85	0.76	21.10	16.19	22.35

Table 4-13. Gender differences in response accuracy rate and response time.

There was no significant difference between the performances of people at different age ranges (for accuracy rate: F(5, 44) = 0.205, p = 0.652, η^2 = 0.002; for response time: F(5, 44) = 0.523, p = 0.471, η^2 = 0.005). However, familiarity with computers before acquiring aphasia seem to have significant impact on how accurate people's interpretations are (F(1, 44) = 17.340, p < 0.001, η^2 =

0.153), but not on perception speed (F(1, 44) = 2.038, p = 0.157, η^2 = 0.021). Figure 4-27 shows that if people had used computer before they had the stroke or brain tumor/injury, their interpreted more visual representations for nouns correctly than those with no computer experience before. Also, it seems that with more exposure to web images, people's perception accuracy of image representations increased.

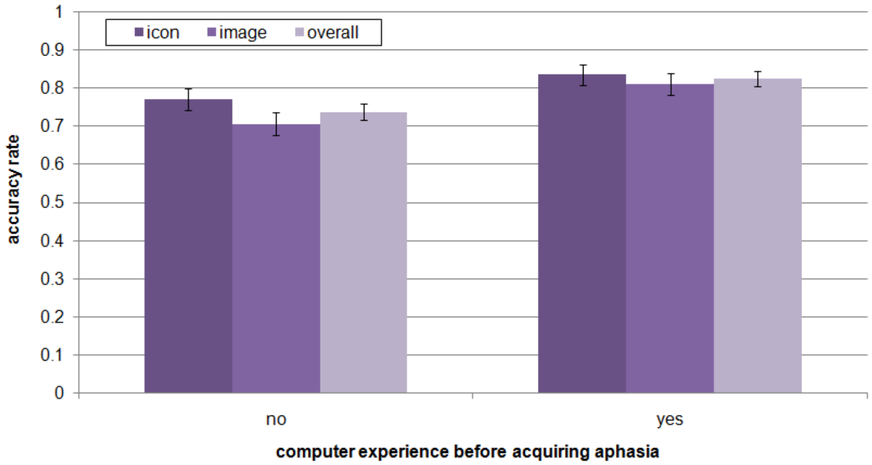


Figure 4-27. Impact of computer experience on response accuracy rate.

4.3.5.5 Cognitive Ability

The cognitive level used in this study was based on the evaluation of the coordinators in the Adler aphasia community center instead of standard metrics. If a participant cannot produce even single words, his cognitive level is said to be low; if a participant can comprehend individual words but not a full sentence, he is said to be median; if a participant can comprehend sentences, he is said to have high cognitive level. In general, people with high and median cognitive level were significantly faster (Figure 4-28 left, F(2, 44) = 18.001, p < 0.001, η^2 = 0.158). In terms of accuracy, people with median and high cognitive level made significantly less error than people with low cognitive level (Figure 4-28 right, F(2, 44) = 11.882, p < 0.001, η^2 = 0.110). It seems aphasic individuals with low to median cognitive abilities were slightly better at icons, but such trend was not found among those with higher cognitive abilities.

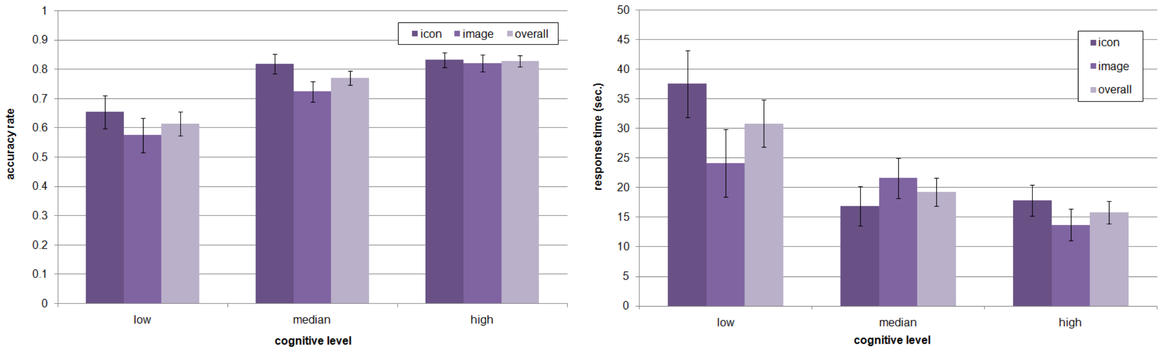


Figure 4-28. Impact of cognitive ability on response accuracy rate (left) and response time (right).

4.3.5.6 Conclusions

To sum up, in the noun naming study with 50 aphasic people, no significant difference between the efficacy as visual representations of web images and that of stylized icons was found, in terms of response accuracy, response diversity, and response time (hypotheses H1, H2, and H3 confirmed). Concrete nouns are significantly easier and faster to interpret than abstract nouns, and linguistic properties such as concreteness and imageability supported this result as well (H4 and H5 confirmed). Post-hoc tests also suggested that people’s computer experience before acquiring aphasia and their cognitive ability after acquiring aphasia had significant impact on the perception of visual representations for nouns regarding to both interpretation accuracy and speed.

4.3.6 Discussion

The results of our experiments indicate that images from the web, despite an uncertain provenance and the lack of uniform conventions, are as effective as icons in conveying meaning.

Vision degeneration due to aging or visual field deficits that often accompany stroke or brain tumor related cognitive impairments influence visual perception as well. Lingraphica's icons are mostly built from simple line drawings in black and white. Compared to images overloaded with varied colors and shapes, they might provide a better and more straightforward representation of concepts for a population more impaired than our test subjects. Using images enhanced via special image processing tools could also be an alternative to compensate for such problems.

Although we provided a magnifying glass to use when viewing the images and icons, the low resolution and poor quality of some of the images probably impacted people's ability to provide us with an optimal guess. Since the images that we used were collected from the Internet, they are of different sizes, shapes, and resolutions; however, icons (ideally) are of identical size and can employ consistent conventions.

One such convention that helped participants understand adjectives represented by icons was deliberate indicators to highlight salient parts of images. For instance, many adjectives are represented alongside their antonym; small is next to big, far is next to near, etc. Specifically, the icon for “little” (meaning small) highlights the smaller of two otherwise identical objects. However, no such cue can be found in the image for “little” that we used (see Figure 4-29). Thus, people came up with a variety of guesses: “small”, “big”, “happy”, “trick”, “circus”, “amuse”, etc. Similarly, for the verb “help”, although both the image and the icon showed an individual reaching out a hand to help someone up, the extraneous information in the scene made it more difficult for subjects to understand the underlying meaning.

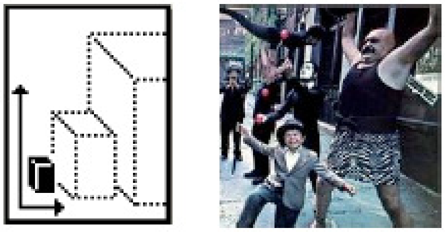


Figure 4-29. Visual representation of the word “little” with an icon (left) and an image (right).

Visual features in addition to human factors affect the results. Figure 4-30 shows the images and icons for the top five most ambiguous nouns in the second study: “appointment”, “government”, “pot”, “trouble”, and “week”. Both images and icons seem to share many of the same failings. First, poor picture quality seemed to affect both images and icons, although more so for images. For instance, the image for “trouble” had poor contrast and appeared too dark, and the icon for “appointment” was also hard to interpret. Second, people sometimes ignored important details. Many participants did not pay any attention to the highlighted part in both the image and icon for “week”, and immediately concluded that it was “month” or “year”. Another problem is that that people focused on only part of the picture. For instance, many participants took the “government” icon as “law” because they saw a balance scale in it. Lastly, sometimes the representation is too far away from the prototypical or common representations. For example, the “pot” image shows a piece of artwork, which is far removed from the planter or cooking pot most people associated with the word “pot”.

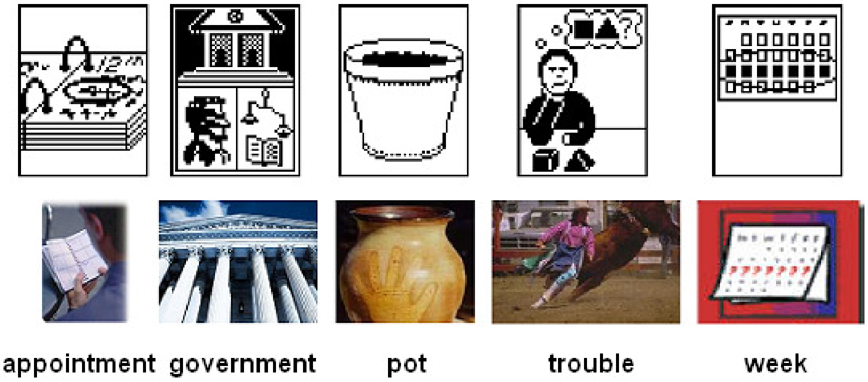


Figure 4-30. Icon and image representations for the top five nouns receiving the greatest number of interpretations.

In the picture perception study with healthy senior citizens, it was discovered that even with context given, some people name a picture with a more generic concept than what was intended while the others came up with a more specific concept than expected, both of which are correct if just look at the picture. For example, three out of twelve people in the first study identified the image for “flower” (Figure 4-31 left) as “rose,” and two out of twelve people put down “dessert” and the other two put down “sundae” for the image representation of “ice cream” (Figure 4-31 right). This phenomenon is typical for concrete nouns, and is an example of the prototype theory mentioned in *Section 2.1.5 Language Storage and Access*. To eliminate the possible confusion caused by the prototype effect, more hints, such as using a combination of pictures as shown in Figure 4-31, should be provided.

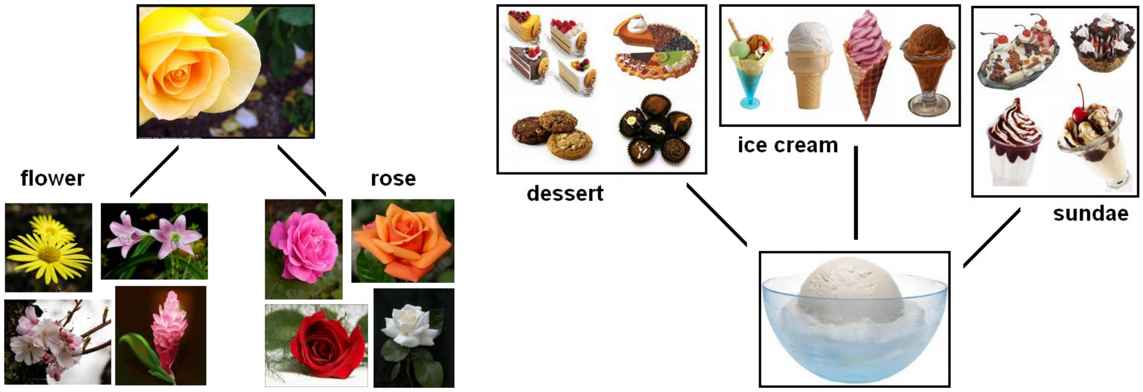


Figure 4-31. Prototype phenomenon in visual representation naming.

Personalized images may further enhance the use of pictures as a language. Because the images from the Peekaboom dataset have been labeled by the consensus of many people, one can expect that people from different backgrounds should be able to retrieve the concept that they depict. This is not always the case, however. In the sentence “We celebrated my wife's birthday in the backyard.”, many participates failed to figure out the word “wife” with the help of the picture given. Their guesses included “parents”, “daughter”, “sister”, “mother”, “girlfriend”, “child”, and “family”.

Despite moving from a free response data collection to a multiple choice test, replacing unimpaired elderly users with sometimes severely disabled individuals with aphasia, and moving away from the context of a sentence, our results are broadly analogous. Despite doubling the number of subjects from the first study and working with a group that is cognitively challenged, we did not find an advantage to using icons over images.

Comparing the two representations, from the designers' point of view, images from the web have great advantages over stylized hand-crafted icons in terms of their ubiquity, low cost, large quantity, and variety. From the users' point of view, not as abstract and generic as icons, images have naturalness of communication. In real conversation, the speaker imparts an implicit context by his age, background, and social status. Taking the “We celebrated my wife's birthday in the backyard.” as an example, even if the viewer has no prior information about the speaker, it will still be easier to correctly interpret the intended meaning if the pictured person is the speaker's actual wife. Another example comes from the sentences we paraphrased from blogs. The original sentence for the paraphrased sentence “My family didn't have money for school.” is about the family having no money for a nephew accepted into Harvard. While it would be difficult and expensive to create icons for every university in the world, users should have the ability to include topics important to them, thus restoring some of the local context lost by the generic vocabulary offered by existing assistive technologies.

4.3.7 Conclusions and Guidelines for Static Visual Representations

The results of the two studies show that web images are as effective as icons in conveying concepts in different parts of speech visually. Both images and icons perform especially well with concrete nouns, but either was as effective with verbs and adjectives.

Various reasons can lead to misinterpretation. Realistic images may provide too many details that distract people, while abstract icons lack details to differentiate similar concepts. Visually perceiving abstract nouns, verbs, and adjectives requires more imagination, association, and evocation from people. Picture quality, impaired vision, possible cognitive degeneration and personality traits may all impact the accuracy and efficiency of the interpretation to a certain degree.

Since this work targets a population with communication impairment, the collection of the source words, the selection of pictures, and the experimental design were carefully tailored. The disability of the study participants was taken into consideration and an aphasia-friendly interface was created. Input and feedback from domain experts who work with this population as well as from their supportive social networks were collected. Exploring the problem space first with participants with similar demographic background to the target population provided useful insight in our working hypothesis, while avoiding all the troubles in recruiting participants from such an isolated community. In addition, piloting with an elderly able population avoids the challenges of communicating with people with aphasia.

A set of guidelines on how to create effective static pictorial representations for AAC vocabularies have been generated based on the observation and results from the two studies. The guidelines are useful for both web image selection and icon creation.

(1) Ensure good picture quality, including appropriate size, brightness, contrast, and color scheme. A lot of the times the intended information is embedded in the details and people need to be able to detect those clues to make correct interpretation. This also takes into consideration the possible vision degeneration of the viewers.

(2) Avoid distraction from the background. Especially for web images, there are often many irrelevant objects or events appearing in the scene, which can cause distraction and

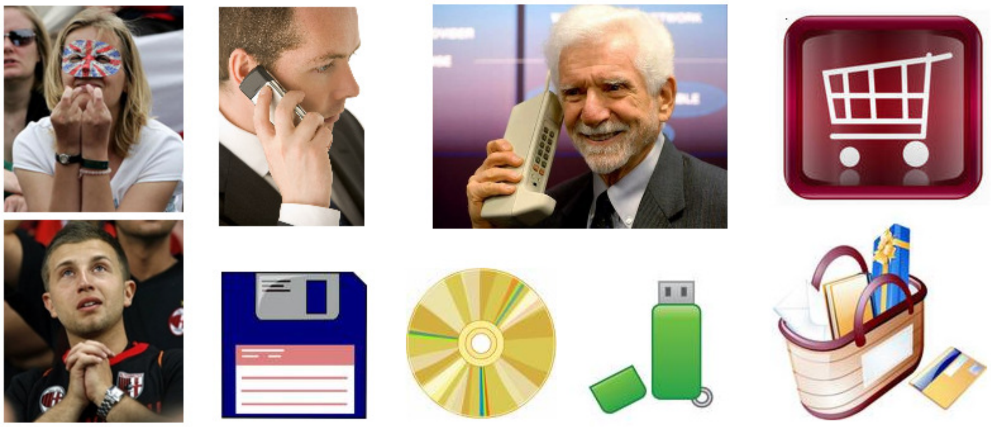
confusion. Therefore, a good representation should stick to a simple background in order to eliminate potential misinterpretation.

(3) Highlight the main content or hints if possible. It can help to direct people's focus to the intended meanings. This is very helpful for conveying adjectives.

(4) Pay attention to prototype effect, especially for representations on concrete nouns.

(5) Choose a more generic scene for representations of abstract concepts, since people easily emphasize on the specific objects, people and events and ignore the big picture.

(6) Choose familiar scenes and be sensitive to familiarity differences introduced by factors like cultural background (Figure 4-32 (a)), education/economic status (Figure 4-32 (d)), and age (Figure 4-32 (b)(c)).



(a) "hope" (b) "cell phone" (top) (c) "disk" or "save" (bottom) (d) "shop/shopping"

Figure 4-32. Examples of representations that are affected by familiarity.

In this chapter, it has been shown that static pictures (both images and icons) are not as effective in conveying verbs. Therefore, new multimedia stimuli as visual carriers for verb concepts were explored, which is introduced in the next chapter.

CHAPTER5 DYNAMIC VISUAL REPRESENTATIONS FOR VERBS

5.1 Difficulty in Visualizing Verbs and Proposed Approaches

5.1.1 Difficulty in Visualizing Verbs

Verbs, a lexical category indicating the presence of a state, existence or operation of an action, are an indispensable part of English speech (Miller1991). In English, although there are more nouns than verbs in the vocabulary, certain verbs are often of higher usage frequency than nouns. For instance, “see,” “give,” and “make” are some of the most commonly used words in English. However, compared to nouns, verbs are more challenging to visualize. As shown in the image vs. icon study (*Static Image* section), the power of pictures in illustrating verbs was significantly lower than in illustrating nouns, which are names for people, places, objects and abstract ideas. According to research in noun and verb naming with object and action pictures with children (Davidoff1995) and people with aphasia (38 publications between 1984 and 2005 summarized in Matzig2009), verb naming from pictures is developed later and verb naming deficiency is more severe than noun naming among aphasic individuals. There could be several reasons for this.

For natural languages (e.g. English, German, Kaluli, Japanese, Mandarin Chinese, and Turkish), the acquisition of verbs is often reported to be later than the acquisition of nouns (e.g. Gentner1978). Many more nouns than verbs are produced and comprehended at the early stage of language learning (Gentner2006). In addition, research on aphasia (e.g. Goodglass1976) revealed that verbs are more vulnerable than nouns to language area impairment in the brain. The Natural Partitions hypothesis (Gentner1982) is one of the theories that attempt to explain this phenomenon. The Natural Partitions hypothesis suggests that (a) the linguistic distinction between nouns and verbs is based on “a preexisting perceptual-conceptual distinction between concrete concepts such as persons or things and predicative concepts of activity, change-of-state, or causal relations;” and (b) noun categories are conceptually simpler or more basic than verb categories.

According to the Natural Partition hypothesis, nouns, especially concrete nouns, have transparent semantic correspondence to the actual world – the world that people perceive through senses. Children have already formed concepts about persons, places, and objects, and learning nouns is basically a process of mapping the words to the concepts. In contrast, the relation between verbs and the perceptual-conceptual world is less transparent. Verbs, even concrete verbs like “run” and “turn,” are more variable in their semantics than nouns. According to WordNet statistics (WordNet2010), on average each verb has 2.17 senses vs. 1.24 for nouns. Within the WordNet lexicon, the ratio of polysemous words and senses to monosemous words and senses is 0.59, whereas the ratio is 3.83 for verbs. People in general have trouble making fine distinction among ambiguous verb concepts. Even adults show significantly better noun-actual world mapping than verb-world mapping (Gillette et al.1999).

On the other hand, nouns (especially concrete nouns) can be organized into a well-structured semantic hierarchy, and noun concepts categorized together usually have intuitive similarity in semantic features. On the contrary, the organization of verbs is much looser and shallower, with less shared semantic properties (e,g. WordNet1998, Huttenlocher1979), which suggests that fewer semantic clues are available to help with verb sense disambiguation. Polysemy and semantic organization differences suggest that the clarity requirement is higher for the non-verbal representations for verbs than those for nouns.

Research has looked into factors that may affect noun and verb naming, and aging is one of them. It has been confirmed by many studies that action naming declines as people age (Ramsay et al.1999)(Mackay2002). This reveals the age-related verb retrieval deficit. It was also shown that aging affects action naming and object naming in the same way (Mackay2002). However, the Action Naming Test (ANT) used in these studies focuses only on action verbs, and use only still pictures are the stimulus. No research has investigated the age-related effect with dynamic visual representations.

The discussions above examine the difficulty in visualizing verbs from the linguistic perspective. However, the visual representations themselves are problematic. Verbs are used to indicate ongoing action, an existing state or the change of condition, which are usually unfolded over time. Ordinary static pictures may fail to portray the temporal information. An ordinary static picture shows a frozen moment, and people and objects in the scene appear to be in a particular position and status. When people look at a picture, they first identify the concrete concepts that are visually revealed. If no hint of a movement or change of condition is available, it is hard to direct people’s attention to generate verbs.

5.1.2 Static vs. Dynamic Pictures for Verbs

Special photo taking and processing techniques have allowed changes and actions being caught in a single static picture (Figure 5-1). Size, position (e.g. the sun in the upper left picture in Figure 5-1), and color (e.g. tree and grass in the upper right picture in Figure 5-1) are information utilized by visual perception to map visual stimuli to concepts appearing in the actual world. If the differences in these visual cues are detected, people can form the idea of “something has changed.” Techniques such as multiple exposure/long exposure or blending/morphing are used to trace the changes and/or reveal the contrast. Similar effects are used to show motions and actions in a single static picture. The “ghost image” (as in the bottom left photo in Figure 5-1) and composite images (bottom right photo in Figure 5-2) of the same subject portray the continuous movement both in position and location. The middle photo at the bottom of Figure 5-1 shows another effect that is common in action shots: the center of focus is the picture is clear while the background gets blurry due to underexposure resulting from the movement of the camera following the main subject. These special effects for capturing actions and motions mimic the persistence of vision phenomenon (see the next section) to show the dynamics developed through time. They basically compress the temporal information into one view. However, the creation of such images requires special skills and/or equipments. They are in general expensive to make, and thus are limited in availability.

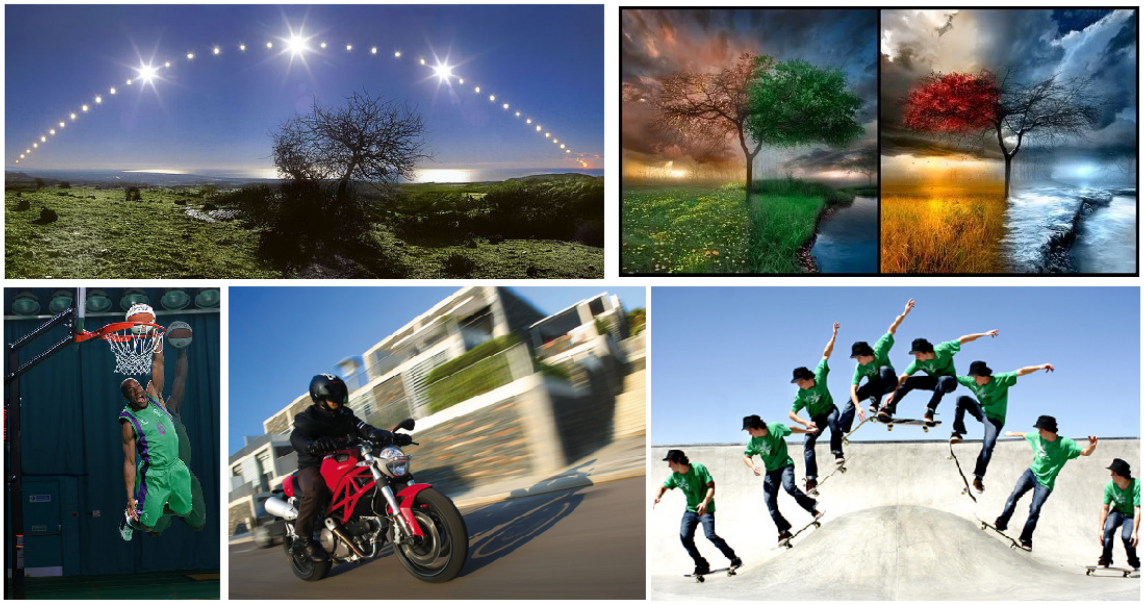
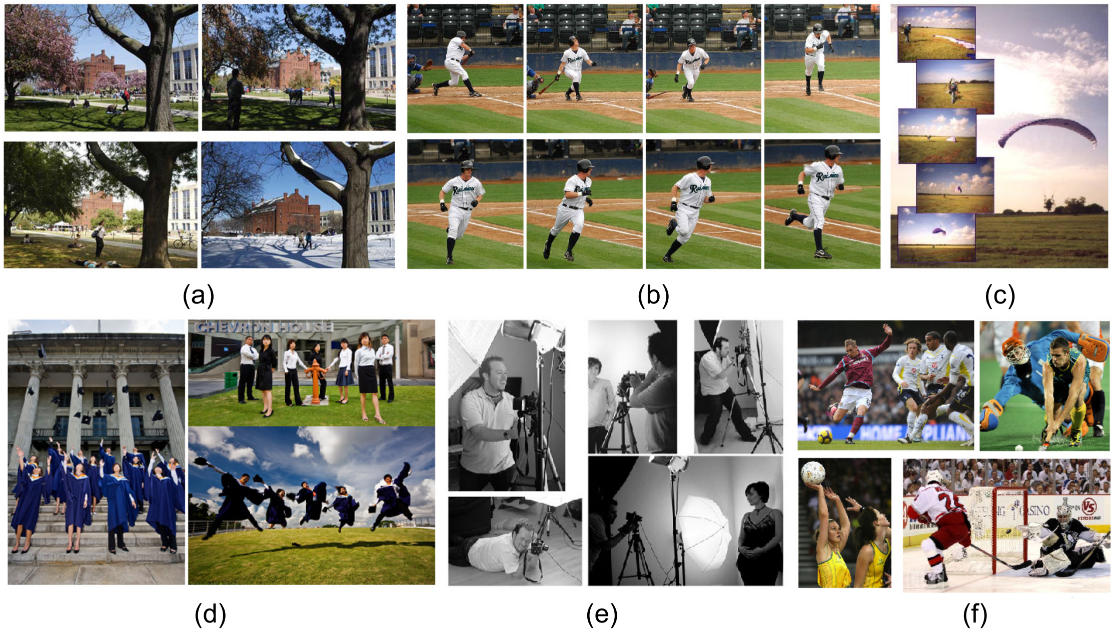


Figure 5-1. Examples of actions and changes in a single static picture.

The special effects in a single static picture can be decomposed into several ordinary static pictures. For example, Figure 5-2(a) presents the change of seasons in four images. Figure 5-2(b) and (c) are action shot sequences on hit and run in baseball and the start of wind surfing. The time interval between shots in (b) is much shorter than in (c). Again, by following the picture sequences, people are able to develop the sense that people/things in the images are in continuous movement/changes. The format of multiple pictures also provides the chance to look at the same event from different angles (Figure 5-2(e) for “shoot: record on photographic film”) or the same action in different context (Figure 5-2(d): “graduate: receive an academic degree upon completion of one's studies;” (f): “shoot: score”).

Using multiple pictures to evoke one concept may have the following benefits. This may be an advantage for depicting more complex and abstract verb concepts. A more complex action usually consists of several simple movements, hand gestures, and/or change of facial expressions. For example, “pick” as “picking an apple” involves stretching the arm, reaching out the hand, touching the apple, twisting and pulling, and then holding the fruit in hand. Verbs for events such as “graduate” involve different actions. The simple movements and actions may vary in different context. If not carefully selected, specific movement may attract more of people’s attention and the bigger picture may get ignored. For example, if just looking at each individual pictures in Figure 5-2(f), people may interpret them as “kick,” “hit,” and “throw.” Presenting multiple pictures of different situations in which the action or event occurs may help to eliminate choices and conclude on verbs at higher level, such as “shoot.” This could be an advantage for distinguishing meanings as well. For example, Figure 5-2(e) and (f) are all representations for the verb “shoot” but are in different senses.

Despite its potential of providing additional cues for verbs, the multiple-pictures stimulus has the risk of introducing confusion with increased information complexity. It is especially the case for multiple pictures showing different contexts (Figure 5-2(f)). Compared to continuous shots (Figure 5-2(b)), the associations or semantic meaning shared across pictures with assorted scenarios may not be as intuitive. Each individual picture has a different set of people, objects and environment for people to digest. It already doubles/triples/quadruples the perception overhead, nevertheless the mental effort required for extracting the commonality among all pictures.


(a) (b) (c)

(d) (e) (f)

Figure 5-2. Examples of changes, actions and events in photo sequences.

Unlike images (photos), graphic icons can not apply those photography techniques to illustrate change in status and actions, but they can resemble the effects (as shown in the first two cartoons on the right column in Figure 5-3). There are other ways to emphasize verb concepts in the iconic representations as well, including (1) lines (solid or dash) for depicting the trace of the action as in the “run” and “throw” cartoon on the right in Figure 5-3; (2) deformation of shapes (the legs of the running chicken); (3) sound (e.g. the alarm clock ringing in the first comics on the left in Figure 5-3 and the “BOOT” and “SPLAT” sound in bottom left comics in Figure 5-3); and (4) thought and speech bubbles (e.g. the top left comic strip in Figure 5-3). In addition, icon strips (sequence of graphic representations) provide key frames of a continuous event, similar to the action shots. People need some imagination to fill in the gap between discrete moments shown in the strip. Usually (as shown in the second and third comics on the left in Figure 5-3), the subject(s) is the same, while the background, the props, as well as the pose, gesture, and facial expression of the character(s) are different as an indication of the progress of the story.

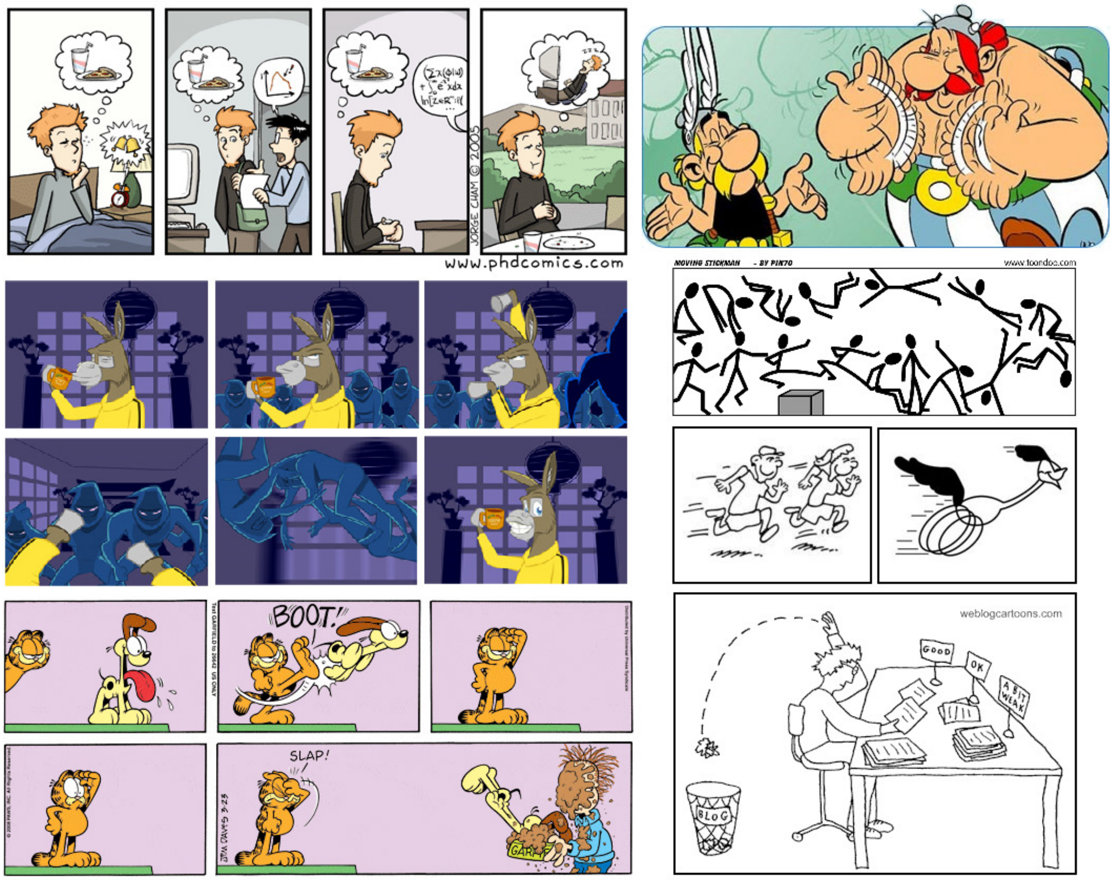


Figure 5-3. Examples of pictorial communication in everyday life: graphs, signs, media and advertisements, arts, cards, maps/flags, comics, etc.

Multiple visual (pictorial or iconic) representations may be able to depict concepts that are usually associated with verbs (e.g. existence, change of status, and actions). However, laying out all the pictures for a single concept takes up a lot of space. Quite a certain amount of information is actually repeated in the picture sequence, and the duplication still consumes memory space and perception power. To overcome the shortcomings while maintaining the positive features in action shots and comic strips, dynamic pictures (i.e. animations: Figure 5-4(a)(b) and videos: Figure 5-4(c)(d)(e)) become a potential candidate for visualizing verbs. Technical speaking, a dynamic picture is a continuous showing of discrete static picture sequence in the same frame. The switch between pictures is so quick that the viewer does not realize it. Dynamic pictures can show simple action such as “push (button)” in Figure 5-4(a) and “spill (water)” in Figure 5-4(f), change of status such as “dissolve” in Figure 5-4(c) and “surprise” in Figure 5-4(e), as well as events like “cook” in Figure 5-4(b) and “arrive” in Figure 5-4(d). Special effects can be applied to the rendering of each frame of the dynamic picture and also the showing process.

Compared to static representations, a dynamic picture can better reinforce the sense of time. It has smoother transit between pictures and takes less space. As flipping through the frames, the part that has been changed gets updated while the consistent background appears to be static in the vision, which saves the effort to reacquire duplicate information. However, since the pictures are not shown at the same time, people need to memorize the initiate status to detect the differences. Also, it is harder for a dynamic picture to present different contexts for the same verb. The discontinuity across frames may simply cause distraction and confusion.

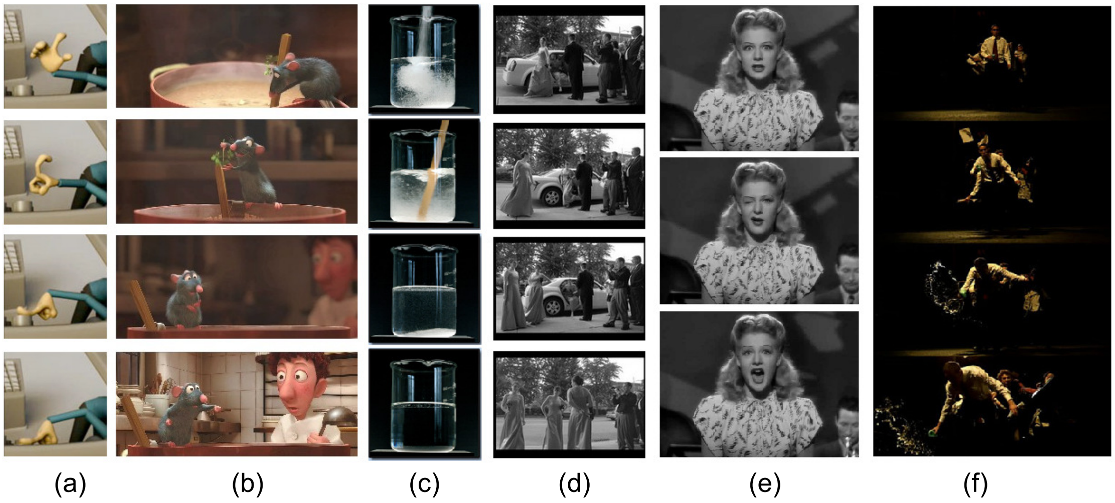


Figure 5-4. Examples of animation and movie frame sequence on actions and events.

5.1.3 Perception of Motion and Perception of Dynamic/Motion Pictures

In the previous chapter, the visual perception of static pictures is reviewed. Here, perception of dynamic pictures (also called motion pictures) is discussed in order to better understand the mechanism that supports visualizing verbs with animations and videos.

As mentioned earlier, when visual signals reach the primary vision area in the brain, the ventral stream identifies the objects while the dorsal stream analyzes the spatial information. As the object moves in the actual environment, the change in its position and location is detected by first-order motion perception (changes in luminance) or second-order motion detection (moving contour defined by qualities like contrast and texture). First- and second-order motion signals are extracted from successive images on the retina, and integrated in the visual area V5/MT in the extrastriate visual cortex to form a global sense of motion of the moving object/surface.

The perception of motion pictures takes a different path, because people form a perceptual illusion of motion without existence of the actual movement. People used to believe that the perception of motion from dynamic pictures come from the persistence of vision phenomenon. From the moment lights reach retina through lens in the eyes, it takes time for the optical signals to be sent to and processed in the brain. Even often the light source gets cut off during the transmission process, the brain still “thinks” that there is an afterimage retained on the retina. This phenomenon of the eye is called persistence of vision (Figure 5-5(a)(b)). An afterimage can last approximately 1/25 of a second, and during this period, a successive image formed by new light sources will be rendered on top of the afterimage (co-existence or combined) (Anderson1993)(Ascher1999). Later, people found that persistence of vision may explain fusion of scenes, but is inadequate to decode the illusion of motion.

Nowadays, another more plausible theory involving two famous perceptual illusions discovered by Max Wertheimer in 1912, i.e. the phi phenomenon and the beta phenomenon, is proposed as reason for motion perception from succession of still images. In the phi phenomenon (Figure 5-5), the succession of static images with omission creates the illusion that a “ghost” image with the background color (a white dot) is jumping along the positions. In the beta phenomenon, the illusion is different – it is the foreground image (the blue dots) that is moving, taking turns filling the blank spot. One explanation for the differences between phi and beta is that, in phi phenomenon, the switch of pictures is so fast that the afterimage (from persistence of vision) of the previous omitted shape still exists, which gets blended with the background color and yields the ghost shape with a different color. Whereas in beta, the switch interval is long enough that the afterimage is no longer retained, and all that are perceived is the foreground objects. More advance than persistence of vision, phi-beta theory suggests that the brain can generate a perceptual illusion of motion by combining optical signal from separate events. This is considered to be one of the basic theories for film making.

All in all, individuals can perceive actions and changes from dynamic pictures, and thus, animations and videos are viable as representations for verbs.

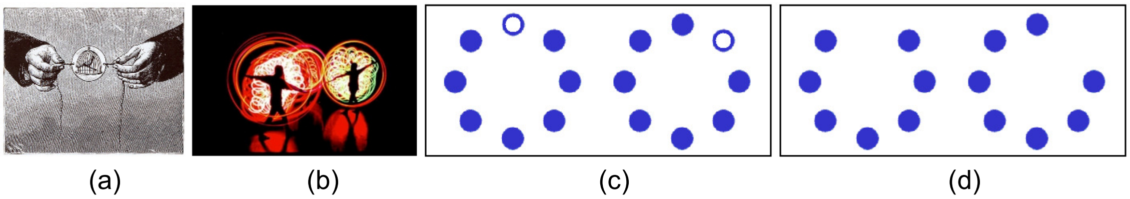


Figure 5-5. Examples of the perception of vision, the phi, and the beta phenomenon.

5.1.4 Visual Verb Representations for Communication Support

Current research on visual stimuli for verbs, including line drawings (Rogers1987), photos (Druks2000), animations (Vanrie2004), and videos (Blank2005), emphasized actions and movements having to do with postures, gestures, and observable manipulation. This small fraction of verbs cannot satisfy the needs of normal communication. In general, there are a lot fewer datasets with pictorial representations for verbs. Two of the main sources are AAC symbol systems (see *Background Work* section) and picture vocabulary for language learning for children. Most of the visual verbs are iconic (Figure 5-6) and only a few provide dynamic representations (e.g. Lingraphica animations, Figure 5-6 bottom).

Besides the small quantity, most of the visual verbs are for concrete actions and events. Very few verification checks on evocativeness have been conducted on the existing visual verb inventory. Evaluation of the efficacy of those visual representations is usually limited to two stimuli, like between icons and animations (Baecker1991). The work was restricted to actions on computer interfaces. There is a lack of studies across all possible visual stimuli, especially with videos. Furthermore, there is a short of guidance on making good valid visual representations for verbs.

My thesis research aims to explore new multimedia stimulus, i.e. videos, as representations for conveying verbs in daily communication. The effectiveness of videos as verb carrier was verified through two studies described in the next section. Guidelines on designing visual verbs were proposed based on observation and analysis of the studies.

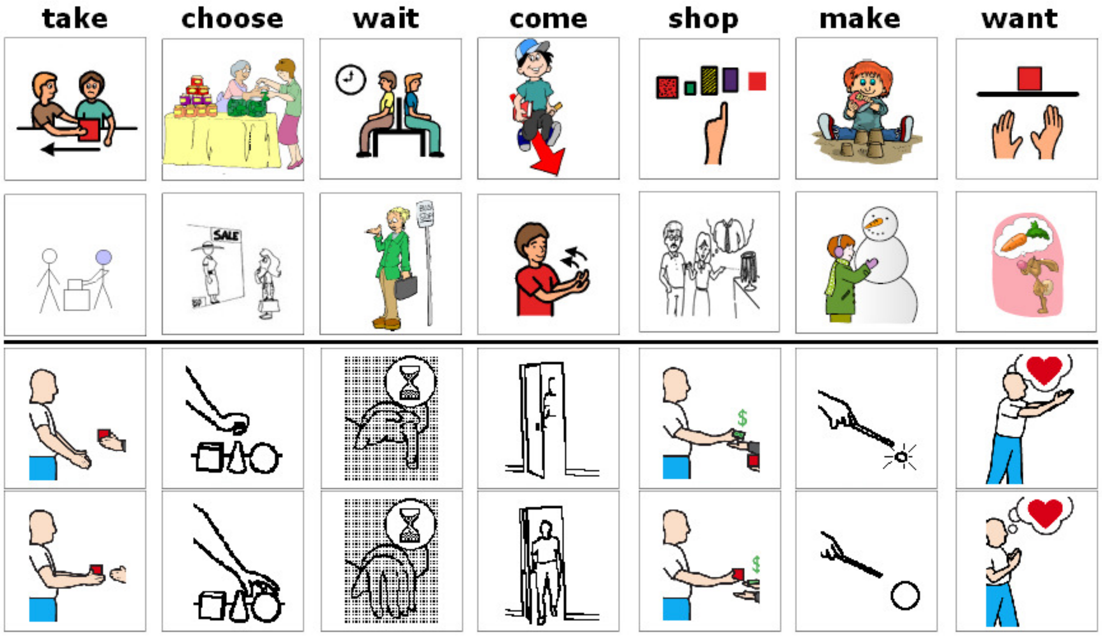


Figure 5-6. Examples of representations for verbs: icons (above) and animations (below).

5.2 Creation and Evaluation of Dynamic Visual Verbs

5.2.1 Research Questions

As mentioned in the discussion in the previous section, verbs in general are harder than nouns to comprehend and produce, to name from pictures, and to present through visual stimuli. Given the essential role that verbs play in communication, providing effective visual representations for common verbs is critical for any AAC vocabularies. Traditionally, single iconic representations (mainly icons and some animations) have been used to depict actions and events. Icons and animations have the advantage of minimizing distraction and providing hints of movement through graphical elements such as lines and symbols. However, as the photography and filming techniques advance, multiple images and videos with good quality have the potential of conveying verbs as well. The evocativeness power of each visual stimulus and the comparison across all are under-investigated. Furthermore, little research has looked into how to visualize abstract verbs which are also important for everyday communication.

In my thesis research, I am trying to address the following research questions:

(1) What visual stimuli are the most effective in conveying verbs? I would like to explore the effectiveness of dynamic vs. static representations.

(2) What kinds of verbs are easy/hard to visualize? I would like study the influence of various linguistic properties on verb naming.

(3) Is there any age-related difference in verb naming with dynamic and static representations? Since the elderly population has a much bigger risk of acquiring aphasia and many of the people experiencing language barriers are senior immigrants, it is important to understand the possible impact aging may have on interpreting visual verbs.

(4) What are the visual cues that enhance/hinder the efficacy of visual verbs? If a person fails to identify the intended verb from a picture, it may be due to his/her own verb retrieval deficit, but also very like, it is because the given representation is badly designed. I would like to come up with design guidelines for creating effective visual representations for verbs that can truly help with comprehension.

To answer these questions, two studies were conducted, comparing the performance of four visual stimuli, i.e. a single static picture, a panel of four static pictures, an animation, and a video, in illustrating commonly used verbs in various domains with both young and old adults. The first study is a non-cued naming test, and the second study is a comprehension test given sentence context.

5.2.2 Visual Verbs Study I (Non-cued Naming)

5.2.2.1 Visual Verbs Study I (Non-cued Naming), Experimental Design

The goal of the individual verb naming study is to determine the effectiveness of four visual stimuli for verbs in nine domains and the possible age-related effect in verb naming. The study used a 4 by 9 by 2 factorial design. The independent variables are visual stimuli (i.e. single static picture, multiple static pictures, animation and video), verb domain (nine categories based on WordNet, see Table 5-1), and age group (young vs. old adults). The dependent variables are response accuracy, response diversity, response time, self-reported difficulty, and self-reported preference. The control variables are participants' vision and word finding ability as well as word frequency. The control made sure that there was no age-related vision degeneration affecting elderly participants' performance, all participants had similar literacy skill, and word frequency, which has been determined to impact the speed of word retrieval, would not be a confound variable.

Several hypotheses were proposed for the study, listed below:

H1. Dynamic pictures are more effective than static pictures in conveying verbs.

H2. Dynamic pictures are more effective than static pictures in conveying verbs for young adults.

H3. Dynamic pictures are more effective than static pictures in conveying verbs for old adults.

H4. The Video mode is the most effective visual stimulus for verbs.

H5. The Video mode is the most effective visual stimulus for verbs for young adults.

H6. The Video mode is the most effective visual stimulus for verbs for old adults.

H7. Verbs in more concrete domains like *contact* and *motion* are easier to interpret.

H8. Performance of young adults in verb naming is better than old adults with all four stimuli.

H9. People consider static representation for verbs harder to interpret than dynamic ones.

H10. People prefer videos for verbs among all visual stimuli.

5.2.2.2 Visual Verbs Study I (Non-cued Naming), Data Preparation

As preparation for the study, I generated a list of 48 common verbs across nine domains, and assigned each verb with four different visual representations.

The tested list consists of 48 uppermost frequently used verbs obtained from the spoken materials such as conversation transcript in the British National Corpus (BNC) (kilgarriff2008). To begin with, I acquired the top 60 most commonly used verbs based on the ranking of verbs sorted in descending order in all forms (definite base form, past tense, and present continuous tense). With a linguist and a speech-language pathologist, I compressed the list by (1) eliminating words that are less common in American English (e.g. “hang” and “reckon”); (2) removing verbs “like” and “know” which usually appear in conjunction phrases “..., like, ...” and “you know;” and (3) keeping only the one with the highest rank given a group of words with similar sense (e.g. “look” and “watch” and “say” and “tell”). The final number came down to 48. The verbs, each assigned with its mostly used sense, can be categorized into nine domains (Table 5-1) according to their functions and association available in the WordNet lexical information.

Domain	#	Verbs
cognition	5	pick, read, remember, think, wonder
communication	5	ask, call, mind, say, thank
consumption	4	eat, make, need, use
contact	6	get, put, shut, sit, take, write
emotion	5	feel, hope, love, want, worry
motion	9	bring, come, go, leave, start, stay, stop, turn, wait
perception	3	hear, look, see
possession	5	buy, find, give, keep, pay
social	6	bet, help, let, play, try, work

Table 5-1. List of 48 verbs categorized by WordNet domains.

The four visual representations we tested were a single static image, a panel of four static images, an animation, and a video clip. The images came from two sources. One is a contest that we set up on Flickr (flickr2008): given the verbs with targeted sense, calling for people to submit photos that they think as the best illustration. The other is public domain images collected from the web. Image search engines such as Google Image Search (googleImage2008) are not as efficient for this purpose, since the search results are retrieved mainly by keyword matching of their surrounding texts instead of the image content itself. Available computer vision databases seldom contain verb classes. Six images per verbs out of all that we gathered were presented to seven raters (graduate students in different fields) to assess their ability to evoke concepts. Summing up all the ratings, we assigned the most preferred image to the single image mode (Figure 5-7(a)), and together with the three next preferred ones to the four images mode (Figure 5-7(b)).

The animations (Figure 5-7(c)) were mostly obtained from Lingraphica by turning its icon strips into animated gif files. There were three cases (Table 5-2) in which I had to recreate new animations. The new creations were consistent with the Lingraphica style. For example, the new “pay” animation was made switching the object in the “give” animation to a dollar sign “$”.

Verbs	Reason
make, mind, pay, use, bet, shut, worry, wonder	no Lingraphica animation available
pick, turn	in a different sense
get, leave	same as “take” and “go” in Lingraphica

Table 5-2. Three Cases Requiring Animations Recreation.

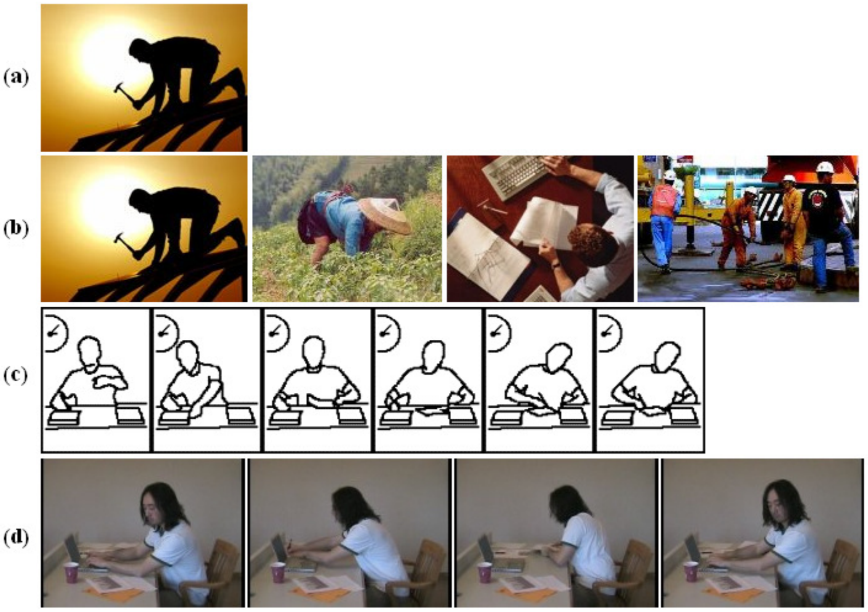


Figure 5-7. Examples of changes, actions and events in photo sequences.

The videos (Figure 5-7(d)) were all filmed by our group. I first looked into other video resources like computer vision databases and YouTube (youtube2008). In computer vision, the data were confined to specific movement such as running and spinning, and at most 20 categories of actions were available. Videos from YouTube are varied, but they are too noise to define and refine as a single verb representation based on the tags. As a result, we shot our own video clips for each verb based on the script selected by four reviewers out of five independently written ones (Figure 5-8). To ensure the quality of the videos as visual representations for verbs, a set of rules were followed during the filming and post-editing process:

(1) Pure white background. It is for minimizing the distraction from the background.

(2) Single leading actor. It is for the consistency of the videos, and more important, for directing viewers' attention to the initiator of the actions and events (Figure 5-8(b)).

(3) Extra hands or feet if interaction required. Transitive verbs may involve more than one person. By only showing the feet and hands of the supporting actor/actress, it is ensure that people get a clear sense of the subject and object of the verb (Figure 5-8(c)).

(4) No sound, text, or lip language. This eliminates the possibility that non-visual cues are embedded (Figure 5-8(d), no actual text was written).

(5) Minimized use of props. Again, it is for avoiding unnecessary distractions and ambiguity.

(6) Clip length three seconds plus or minus one-half second. Through a set of test, our group determined that three second is long enough to convey a verb concept and yet not too long to have information overflow or cause viewers to lose attention.

(7) Applying special effect if necessary, for example, fast forwarding the “make” and “work” video. Special effects can help emphasize on several key frames within a long sequence of action/event.

The scripting and filming of the videos were supported by members of Princeton Sound Lab and Princeton Aphasia Project.

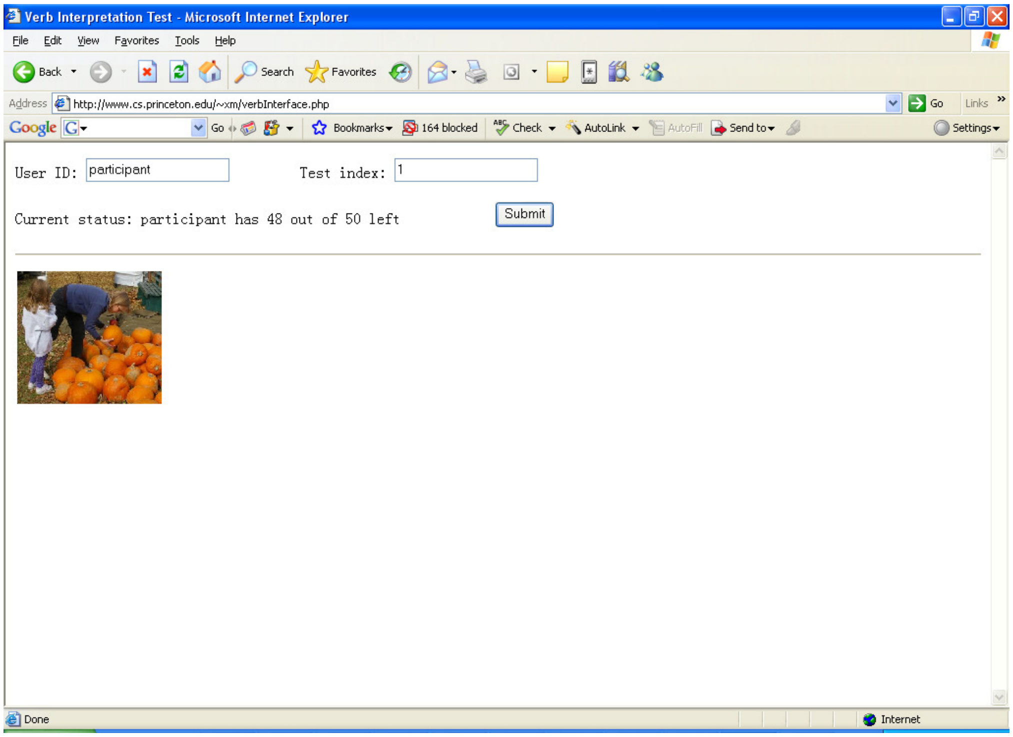
Figure 5-8. Video filming.

5.2.2.3 Visual Verbs Study I (Non-cued Naming), Study Procedure and Participants

The 48 verbs were divided into four blocks evenly so as to align with the four visual stimuli according to the Latin Square design. Verbs from the nine domains were equally and randomly distributed within the four blocks. Verbs in different domains were evenly distributed into the blocks in random orders. When presented to the participants, the sequence of the blocks was also determined by a Latin Square for counter-balancing. The verbs were shuffled within each block.

All visual representations were unified to a height of 132 pixels, and displayed on a web-based interface (Figure 5-9) one at a time. The participants were asked to name the first verb that comes up to their mind when seeing the visual content, and then justify their response in a few sentences. The justification helped to determine the meaning of the response. For example, one participant said “pick, because the woman is picking up a pumpkin from the ground,” and she meant “pick up: take and lift upward.” Another participant said “pick, the mother is picking the best pumpkin out of a pile,” meaning “pick: select carefully from a group.” Responses were noted by the investigator and the response time was recorded at the interface backend. At the end of the study, participants were requested to rank the four visual representations by (1) the difficulty in interpreting to a verb; (2) the speed of coming up with some thought; (3) the confidence in the response; and (4) their personal preference.

This study was conducted in the University of British Columbia, BC, Canada, under the supervision of Professor Joanna McGrenere (Computer Science), Professor Peter Graf (Psychology), and Dr. Barbara Purves (Speech Language Pathologist).



http://www.cs.princeton.edu/~xm/verbInterface.php

User ID: participant   Test index: 1

Current status: participant has 48 out of 50 left

Figure 5-9. Web interface for the individual verb naming study.

The individual verb naming study was carried out with a group of younger participants (20-39) and a group of older participants (55+), 16 people each. The study was located in the Human Computer Interaction Usability Lab in the University of British Columbia, BC, Canada. Participants were recruited via advertisement on the craigslist website and a local newspaper. Besides their demographic information, participants' word finding ability was assessed through the F-A-S Word Naming Test (benton77) and their eyesight (no worse than 20/40) were tested (Table 5-3).

	Age		**Education (years)**		**F-A-S Score**	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
young	30.94	6.24	15.13	1.45	42.19	8.89
old	71.88	5.80	14.38	2.73	40.75	9.23

Table 5-3. Participants' demographic information and word finding ability.

5.2.2.4 Visual Verbs Study I (Non-cued Naming), Evaluation Metrics

To quantitatively analyze the accuracy and distribution of the responses, four metrics measuring different aspects were proposed: correctness, irrelevance, WordNet score, and response diversity. *Correctness* and *irrelevance* are defined as the number of the exact match and the number of totally irrelevant responses separately. Exact match means the response is the same verb in the same sense as what was pre-assigned to the visual representation, proved by the justification. For instance, the verb "pick" given in the study is in the sense of "select carefully from a group."

The response "pick, because the lady is picking pumpkins in the field" is considered as a false response, since it uses the sense "look for and gather." It is actually an irrelevant answer.

WordNet Score (weighed) assess the responses in a six-point scale (Table 5-4) by their WordNet semantic distance to the intended verbs for each visual representation. Responses receive point 6 are exact match while those with point 1 are irrelevant. Taking into account the fact that mainly verbs in the same synset can accurately deliver the message in real communication, we also calculated the weighed WordNet score: the exact match weighed by two, synonyms weighed by one, and the rest zero.

Response Diversity shows the number of difference responses received for each verb under each visual mode, which tells how well the interpretation converges. If the responses spread out in the semantic network, it means the visual representation failed to illustrate the intended verb. In another case, if the responses gathered around a verbs that is different from the target, it means the representation successfully convey a concept, though the wrong one.

Response Time was the interval between the visual stimuli showed up on the interface and the time the participant clicked the button to proceed to the next representation. Although the response time could be influenced by the speed of verb retrieval, the length of justification, the speed of talking, and the speed of interface interaction, it still reflects the difficulty in identifying a verb from the given visual representation.

	case	**example**
6	exactly the same verb in the same sense	buy (obtain by purchase)
5	other words in the same synset as pre-assigned	purchase (obtain by purchase)
4	direct hypernym or hyponym of the pre-assigned synset	pay (give money in exchange)
3	indirect hypernym or hyponym within 4 steps away	shop (do one's shopping)
2	connected by entailment or sister/nephew terms	exchange (give and receive)
1	not connected, or further away than score 3 and 2	donate (give to a charity)

Table 5-4. WordNet score in six-point scale.

	correctness	irrelevance	# of responses	response time
total	0.9863	-0.5829	-0.6856	-0.2656
young	0.9830	-0.5504	-0.6966	-0.1874
old	0.9894	-0.6156	-0.6704	-0.0397

Table 5-5. Correlation between WordNet score and number of matched responses, irrelevant responses, and number of different responses.

The subjective evaluation measures, including naming difficulty, naming speed, confidence, and preference were calculated by counting the number of votes for each stimulus at each rank.

5.2.2.5 Visual Verbs Study I (Non-cued Naming), Results and Analysis

Visual Stimuli and Age Group

In the study on interpreting individual verbs, multivariate ANOVA was applied to analyze the correctness and irrelevance of the responses, the median of response time, as well as the influence of aging. Results showed that the effectiveness of illustrating verbs is significantly different among the four visual representations (Table 5-6). Video mode significantly outperformed the other modes on interpretation correctness (Figure 5-10 bottom); single static image mode was significantly the worst given the number of correct and irrelevant responses (Figure 5-10 top) it generated, even though it took the least time (Figure 5-11) for people to perceive; the panel of four images mode and the animation mode performed similarly in between.

Efficacy of Visual Stimuli					
measures	**df**	**N**	**F**	**p-value**	**η^2**
correctness	3	90	8.910	<0.001	0.229
irrelevance	3	90	7.407	<0.001	0.198
response time	3	90	12.685	<0.001	0.297
Age-related Effects					
measures	**df**	**N**	**F**	**p-value**	**η^2**
correctness	1	30	5.135	0.031	0.146
irrelevance	1	30	6.100	0.019	0.169
response time	1	30	10.787	0.003	0.264
Interaction between Visual Stimuli and Age					
measures	**df**	**N**	**F**	**p-value**	**η^2**
correctness	3	90	1.760	0.161	0.055
irrelevance	3	90	0.929	0.430	0.030
response time	3	90	2.292	0.084	0.071

Table 5-6. MANOVA results on performance of various visual stimuli, age-related effects, and their interaction.

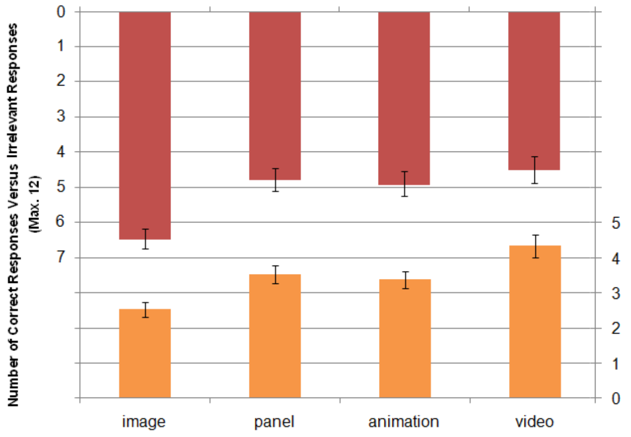


Figure 5-10. Comparison of the numbers of correct (bottom) and irrelevant (top) responses with each visual stimulus.

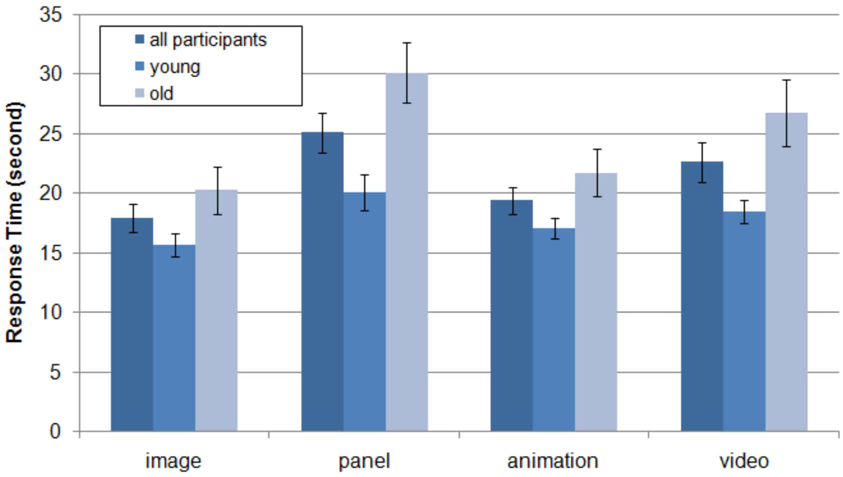


Figure 5-11.Comparison of response time with each visual stimuli.

Comparing the overall interpretation quality between the two age groups by ANOVA, younger participants did a significantly better job than the elderly, with higher accuracy (Figure 5-12), fewer inappropriate responses, and shorter response time (Figure 5-11). Except for the video stimuli, elderly participants had significantly lower WordNet scores than young participants on their responses to single image, multiple images, and animations, regardless of that the trend of differences in performance across the four visual stimuli is similar for young and old adults (Table 5-6). To be more specific, according to response accuracy, the video stimuli stood out significantly for elderly participants, while for younger participants all the multi-picture modes behaved similarly well and significantly better than the single image mode. Young participants' perception time was significantly shorter than that of the older participants' on each mode, and both age groups spent significantly less time on interpreting single static images. The variance in response time was bigger for old adults than for young adults.

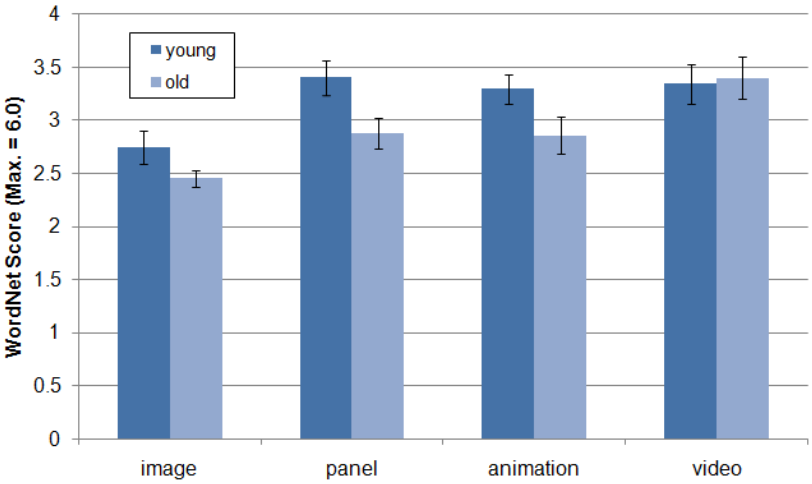


Figure 5-12. Comparison of naming accuracy in WordNet score.

Verb Domains and Other Linguistic Properties

In general, there was no significant difference in the naming accuracy of verbs in different domains (F(8, 78) = 0.784, p = 0.618, η^2 = 0.074). However, different visual representations are significantly different in the domains of verbs that they are good at illustrating based on ANOVA results of weighed WordNet score (F(24, 234) = 2.252, p = 0.001, η^2 = 0.188), though the impact is similar for young and older adults (F(24, 234) = 0.441, p = 0.990, η^2 = 0.043). In general (Figure 5-13), contact, social, and cognition verbs, which are usually associated with certain gestures, are significantly easier to interpret than communication verbs. For contact verbs, video mode performs significantly better than a single static image mode and a panel of four static images mode; for motion verbs, video modes significantly outperforms all three other modes; for the rest of the domains, there is no significant difference among modes, among which, the distraction from facial expressions might be one of the possible reasons why video did not perform as well in emotion and perceptions verbs.

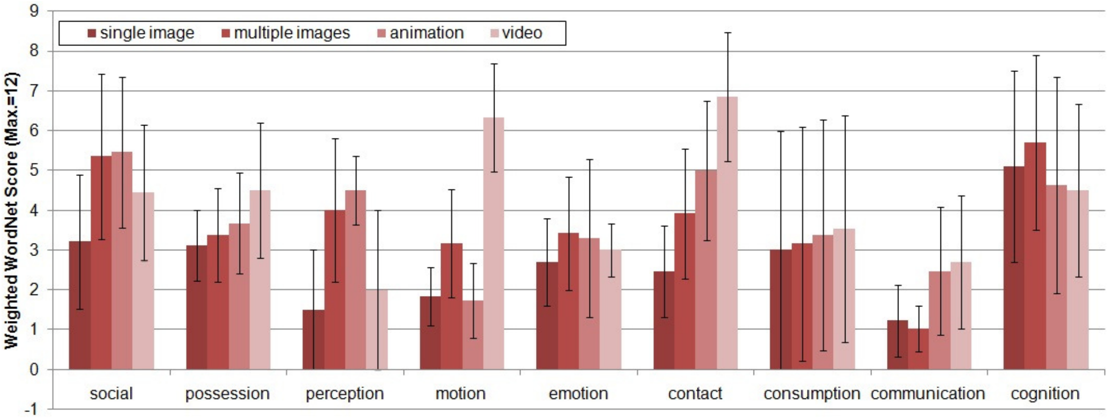


Figure 5-13. Comparison of performance of the four visual stimuli with different verb domains.

In order to further explore what linguistic properties leads to the variation in naming performance on verbs from different domains, post-hoc analysis was conducted to examine the influence of concreteness and imageability of the concepts, and polysemy of the words. Based on MRC Psycholinguistic Database (MRC2008), I divided concreteness (CNC) and imageability (IMG) into three levels respectively: abstract/non-imageable (with scores less or equal to mean minus one standard deviation), median (with scores between mean and mean minus one standard deviation), and concrete/imageable (with scores greater or equal to mean). Results (Table 5-7) showed that only imageability revealed a significant influence on the interpretation. The ability to illustrate verbs with different imageability varied significantly among the four visual modes (Figure 5-14). All the modes worked equally well on expressing highly imageable verbs, however, only the video mode stood out coming to verbs with medium to low imageability. No obvious age effect was found interacting with imageability. For highly imageable verbs, multiple static images were slightly easier to interpret.

Polysemy is measured by the number of senses each word has based on WordNet. ANOVA results showed that polysemy affects the verb naming accuracy (WordNet score) of both young and old participants with a large-size effect (for young adults: F(4, 17) = 1.050, p = 0.408, η^2 = 0.181; for old adults: F(4, 17) = 1.227, p = 0.333, η^2 = 0.205).

linguistic property	df	N	F	p-value	η^2
CNC	2	82	2.063	0.134	0.048
CNC × visual stimuli	6	246	1.800	0.100	0.042
CNC × age	6	246	0.309	0.932	0.007
IMG	2	82	6.982	**0.002**	0.146
IMG × visual stimuli	6	246	2.448	**0.026**	0.056
IMG × age	6	246	0.671	0.673	0.016

Table 5-7. ANOVA results on concreteness (CNC) and imageability (IMG).

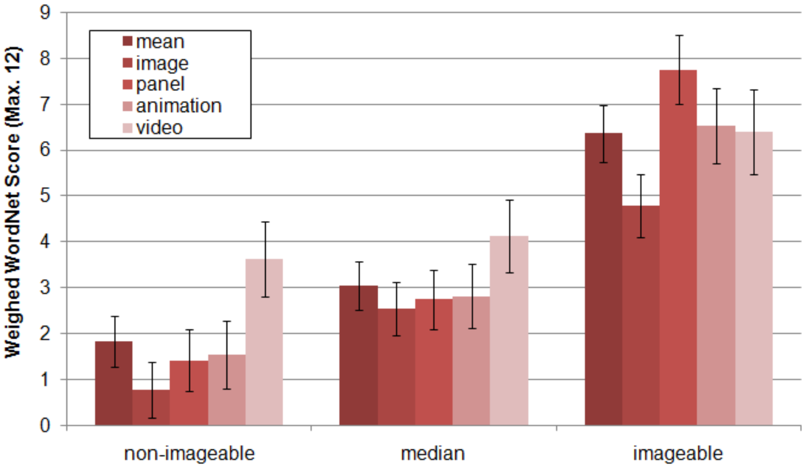


Figure 5-14. Comparison of various visual verb stimuli with different imageability.

Impact of Visual Cues

In order to locate the features employed by each visual stimuli that enhance or hinder the efficacy of the representations in order to generate design guidelines, three kinds of visual cues, i.e. symbol, gesture, and facial expression were examined (Table 5-8), and their effects on people's visual verb interpretation reflected in response justifications were studied (Figure 5-15).

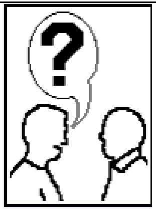
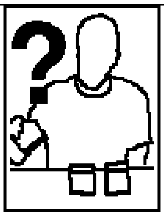
symbol		**gesture**	**facial expression**
direct	**indirect**	waving goodbye for "leave"	irritated look for "mind"
? for "ask" ♥ for "love"	? for "find" ♥ for "want"		
			

Table 5-8. Three visual cues: symbol, gesture, and facial expression.

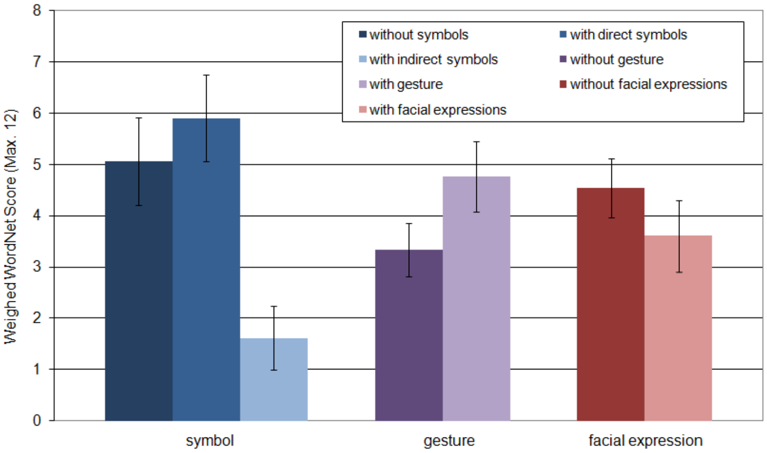


Figure 5-15. Influence of three visual cues.

visual cue	df	N	F	p-value	η^2
symbol	2	74	9.945	<0.001	0.502
symbol × visual stimuli	6	222	4.453	<0.001	0.107
gesture	1	74	1.051	0.309	0.014
gesture × visual stimuli	3	222	1.739	0.160	0.023
facial expression	1	74	0.338	0.563	0.005
facial expression × visual stimuli	3	222	2.508	0.060	0.033

Table 5-9. ANOVA results on symbol, gesture, and facial expression.

Symbols (graphic symbols in the animations as well as the symbolic objects appearing in the image and videos such as light bulb for "remember") were shown by ANOVA results to have strong influence (Table 5-9) on the interpretation, and the animation mode was affected the most (F(6, 111) = 3.152, p = 0.007, $\eta^2 = 0.146$). Figure 5-16 shows that although symbols can improve the perception to some degree if straightforwardly associated with the concept, they cause more confusion when the association is not as clear. Several people interpreted the "want" animation as "love" due to the ♥. It suggests that the use of symbol should be more careful, and icons and animations may not be the best choice to convey abstract concepts.

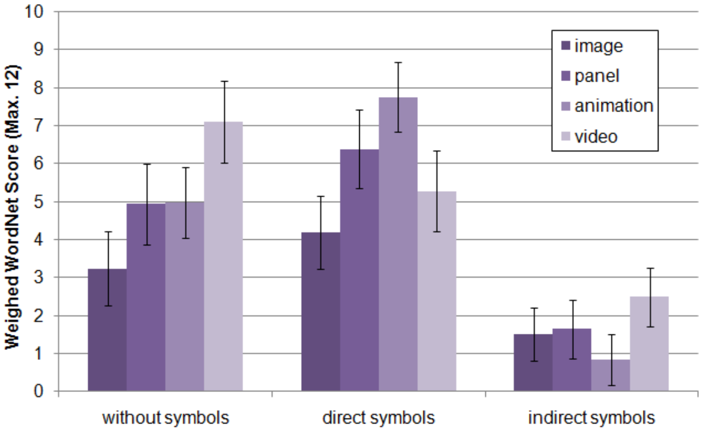


Figure 5-16. Impact of symbol on the four visual stimuli.

Gestures (in images, animations, and videos) seemed to have a positive though not significant effect on the interpretation (Figure 5-17) Familiar poses can evoke the concept relatively easily, for instances the classic thinker pose. However, designers should keep an eye on the possible culture differences in the usage of gestures. For example, people in China have a different gesture for "hope" than crossing fingers. The interaction between gesture and visual stimuli is significant (Table 5-9). It suggests that action and movement verbs are comparatively simpler to illustrate and interpret, especially with dynamic visual representations like videos.

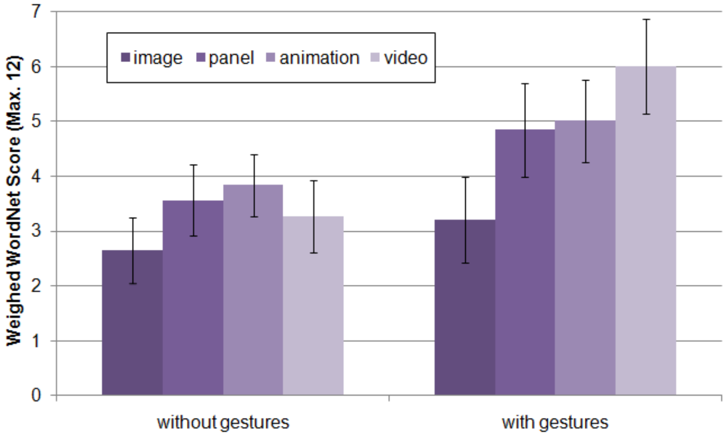


Figure 5-17. Impact of gesture on the four visual stimuli.

Facial expressions of the actor in the videos were said to be more distracted than those in the images, and thus the video mode performed slightly worse than the other modes when the look on the face was shown (Figure 5-18). On the contrary, animations in which the characters were drawn in simple lines with little facial details received no impact. The slight difference was reflected in the interaction between facial expression and visual stimuli (Table 5-9). It suggests that omitting the face of the actor's in the videos for non-emotion and non-perception verbs should be added to the filming rules.

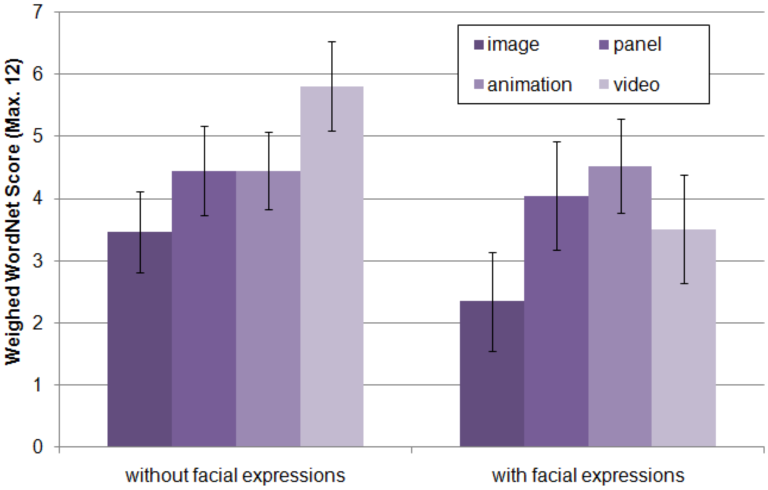


Figure 5-18. Impact of facial expression on the four visual stimuli.

Subjective Evaluation: Difficulty, Speed, Confidence, and Preference

Based on the answers to the questionnaires, participants' feedback on the four tested visual representations from various aspects was summarized in Table 5-10 and 5-11, including difficulty in naming, response speed (meaning how soon people start to have some thoughts, not necessary the final verb, popping up), confidence in their answers, and preference of visual stimuli. It seems older people find moving pictures (animation and video) the easiest to intrigue thoughts and convey to a concept, while four images works as well for younger people. Both younger and older people prefer videos and dislike single static image. People who like or dislike animation and four images are about 50%.

age	stimuli	difficulty		speed		confidence	
		easiest	hardest	fastest	slowest	most	least
all	single image	3	7	6	7	3	11
	multiple images	8	12	6	13	12	8
	animation	10	9	9	8	11	7
	video	12	4	13	4	9	4
young	single image	2	5	5	3	3	6
	multiple images	6	5	4	5	8	3
	animation	5	5	3	6	6	5
	video	5	1	8	2	5	2
old	single image	1	2	1	4	0	5
	multiple images	2	7	2	8	4	5
	animation	5	4	6	2	5	2
	video	7	3	5	2	4	2

Table 5-10. Number of participants commenting on the difficulty, speed, and confidence of visual verb naming.

age	stimuli	preference			
		most preferred	second most	second least	least preferred
all	single image	3	7	14	8
	multiple images	9	4	7	12
	animation	11	4	9	8
	video	9	17	2	4
young	single image	2	2	8	4
	multiple images	6	2	3	5
	animation	5	2	4	5
	video	3	10	1	2
old	single image	1	5	6	4
	multiple images	3	2	4	7
	animation	6	2	5	3
	video	6	7	1	2

Table 5-11. Number of participants commenting on the preference on visual stimuli.

5.2.2.6 Visual Verbs Study I (Non-cued Naming), Conclusion

In summary, the individual verb naming study showed that dynamic pictures were better visual representations for verbs than static ones (hypothesis H1, H2, and H3 confirmed). An age different was revealed in the naming performance (young participants were significantly superior in accuracy and response time; hypotheses H8 confirmed). The video stimulus was the most effective in visualizing verbs for old adults (H4, H5 confirmed). For young adults, any stimulus with multiple pictures or frames had similar performance (H6 partially confirmed). Overall, no significant naming accuracy difference was found across nine verb semantic domains. However, the evocativeness of different visual stimuli on the assorted verb domains was significantly different. The video stimulus significantly outperformed the others on contact and motion verbs (H7 partially confirmed). Highly imageable and less polysemous verbs in general were easier to name from a visual representation. The video stimulus was better than single image, multiple images, and animation on conveying low to median imageable verbs. Visual cues like symbols (positive or negative), gestures (positive), and facial expressions (negative) had influence on the efficacy of the visual stimuli. Based on the post-study questionnaire, people found multiple pictures the hardest and slowest to interpret, and videos the easiest and fastest to interpret. Young and old participants had different attitude and perception of the visual stimuli. Young people preferred multiple pictures and animations, while elderly people preferred the animations and videos (H10 partially confirmed).

5.2.3 Visual Verbs Study II (Cued Comprehension)

5.2.3.1 Visual Verbs Study II (Cued Comprehension), Experimental Design

As discussed above, there are many factors that can lead to misinterpretation if only a single visual representation is shown. Before looking into specific ways to modify and improve those visual presentations to promote the accuracy of individual verb perception, I would like to explore whether the communication context would help people to understand what the picture or video tries to convey. A second study was designed with verbs given in sentences, and participants were asked to interpret the whole sentence. Again, I would like to determine whether verb semantic property and aging have an effect on people’s comprehension of visual verbs in context. The verb in context comprehension study used a 5 by 9 by 2 factorial design. The independent variables are visual stimuli (i.e. single static picture, multiple static pictures, animation, video, and a baseline stimulus which is leaving the tested verb blank in the sentences), verb domain (nine categories based on WordNet, see Table 5-1), and age group (young vs. old adults). The dependent variables are response accuracy, response diversity, response time, self-reported difficulty, and self-reported preference. The control variables are participants’ vision and word finding ability as well as word frequency. The control made sure that there was no age-related vision degeneration affecting elderly participants' performance, all participants had similar literacy skill, and word frequency, which has been determined to impact the speed of word retrieval, would not be a confound variable.

New hypotheses based on the first naming study were proposed for comprehension study:

H1. The Video mode is the most effective visual stimulus for verbs for both young and old adults.
H2. The response time for video is the shortest and the response time for multiple images is the longest.
H3. Interpretation accuracy is significantly better with context than without context.
H4. The performance of various visual stimuli on different domains of verbs is different with and without context.
H5. Symbols, gestures, and facial expressions have the same impact on visual verb interpretation with and without context.
H6. Context properties such as complexity and interpretation accuracy on other parts of speech have a significant impact on the visual verb comprehension.
H7. People prefer dynamic pictures than static pictures as the representations for verbs in context.

5.2.3.2 Visual Verbs Study II (Cued Comprehension), Data Preparation

The same set of 48 most frequently used verbs were tested in the second study. They were embedded in 65 sentences (see Appendix A) collected from blogs of senior citizens in the Ageless Project (ageless2008). The procedure was as follows. First, sentences containing 48 target verbs (with pre-assigned senses) were crawled from the blogs. Then the sentences were simplified by removing complicated clauses, switching less common nouns with frequently used and yet easy to present ones, and paraphrasing. Each sentence has one to three verbs to test, and each verb appear twice among all the sentences.

Based on the block design and Latin Square assignment, the sentences were divided into five blocks evenly (13 sentences each). Verbs in each block were represented by one of the four visual stimuli as in the first study. In addition, a fifth stimulus was added with the verbs left blank so that the participants have to guess based purely on the context, which served as a baseline. The blank mode evaluated how much people can make out merely based on other words in the sentences. Nouns and adjectives in the sentences are illustrated by single image manually selected from the web. One may argue that the mix of text and pictures may cause confusion. However, this was designed for simulating the visual communication to the biggest extends. Function words are generally hard to illustrate and thus were left as texts. Within each trial, the assignment of representation stimuli to block and the order of the blocks were counterbalanced using Latin Square Design. Trials were randomly picked for each participant, and the order of sentences within each block was shuffled for each trial.

5.2.3.3 Visual Verbs Study II (Cued Comprehension), Study Procedure and Participants

The visualized sentences were displayed one after another on a web-based interface (Figure 5-19). Individual representation was uniformed to the height of 132 pixels. The four-image stimulus was displayed as a two by two grid. Participants were asked to interpret the whole sentence word by word and type the entire response in the text field on the interface. When submitted, the response and the time were recorded automatically by the script for further analysis. The same questionnaire was presented to the participants after the cued comprehension study to collect feedback on interpretation difficulty, speed, confidence, and personal preference of stimuli.

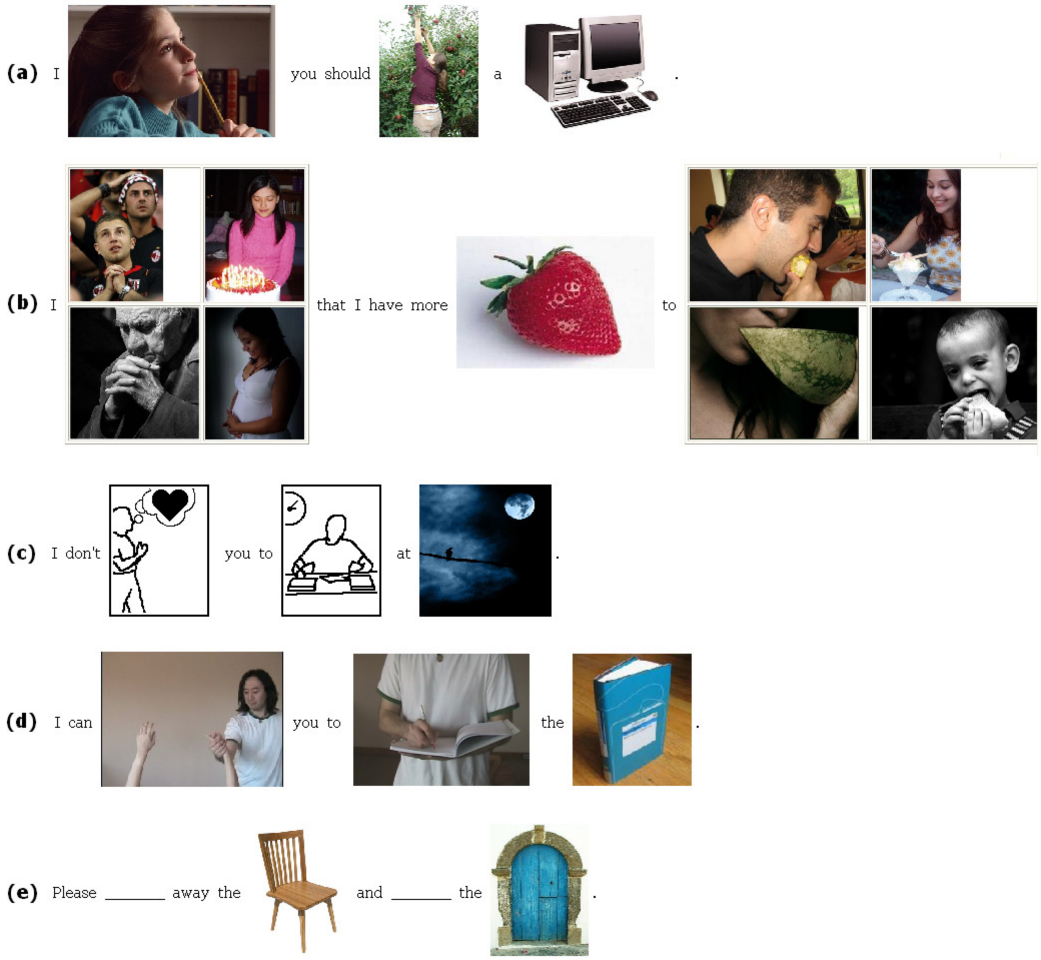

(a) I you should a .
(b) I that I have more to
(c) I don't you to at .
(d) I can you to the .
(e) Please _______ away the and _______ the .

Figure 5-19. Visualized sentences on the web interface for the verb in context comprehension study: (a) single image, (b) multiple images, (c) animation, (d) video, and (e) baseline.

The study was carried out with a group of 25 young participants (of the age between 20 and 39) and 25 elderly participants (55+) who were recruited from the Princeton University student body and a local senior citizen center respectively. The participants' age and their F-A-S Word Naming Test scores are summarized in the Table 5-12. Participants’ eyesight was no worse than 20/40.

	Age		F-A-S Score	
	Mean	S.D.	Mean	S.D.
young	22.60	3.52	49.52	13.74
old	72.88	5.75	44.18	13.88

Table 5-12. Participants' demographic information and word finding ability.

5.2.3.4 Visual Verbs Study II (Cued Comprehension), Results and Analysis

The main evaluation metrics used in the second study were WordNet score (comprehension accuracy) and response time, since the correlation between WordNet score and number of different responses (response diversity) was high (-0.8803). Response time in the comprehension in context study referred to the interval between the sentence showing up on the interface and response submit button being pressed. The original response time measured the time it took to interpret a whole sentence, and could be influenced by speed of typing. In order to get a rough idea of indentifying individual verb might take, the original response time for each sentence was divided by the number of tested verbs. The analysis was conducted by multivariate ANOVA.

Visual Stimuli and Age-related Effect

ANOVA results showed that there is a significant aging effect on interpreting visual verbs (F(1,94)=4.499, p=0.037, η^2 =0.046). Younger participants captured the concepts better than elderly participants based on each visual mode. The visual modes performed significantly differently (F(3,210) = 7.411, $p < 0.001$, η^2 = 0.096), with video mode having a significant better performance over the single static image (Figure 5-20, 5-22). The comprehension accuracy with each stimuli was affected by age in a similar way (F(4,94) = 1.478, p = 0.204, η^2 = 0.016).

Age-related differences were revealed in response time as well (Figure 5-21). It took significantly shorter time for young participants to interpret visual verbs in context than older participants (F(1,646) = 85.926, $p < 0.001$, η^2 = 0.116). The average time required to comprehend verbs in each visual stimuli was significantly different (F(1,646) = 7.317, $p < 0.007$, η^2 = 0.010), and this difference was dissimilar for young and old adults (F(1,646) = 4.586, $p < 0.033$, η^2 = 0.006). To be more specific, it took roughly the same amount of time for young participants to interpret sentences in various visual stimuli. On the contrary, old adults spent significantly more time trying to figure out the meanings of multiple pictures, and they were quicker at animations compared to other visual representations.

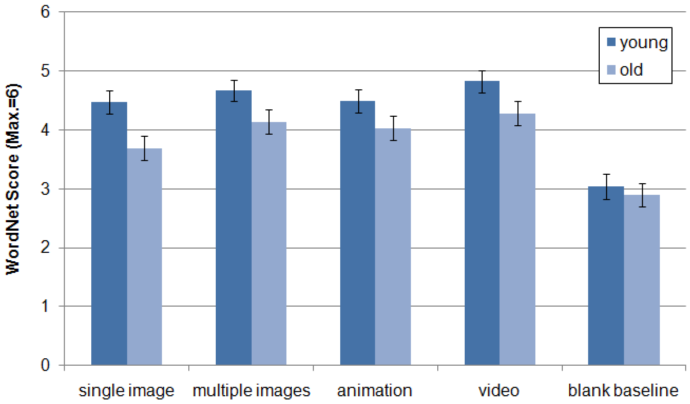


Figure 5-20. Comparison of age and visual stimuli differences in comprehension accuracy.

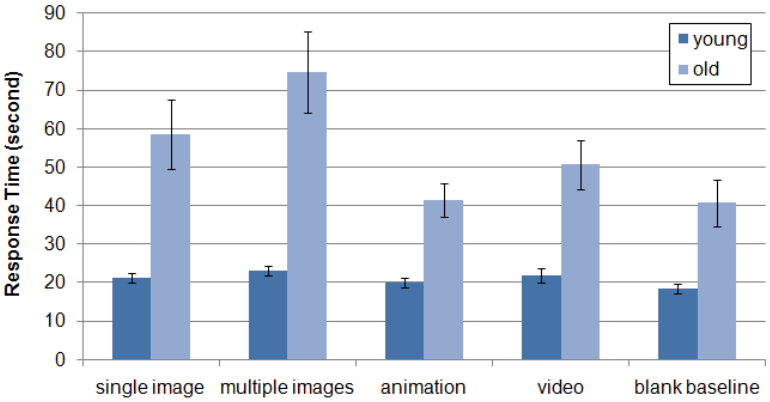


Figure 5-21. Comparison of age and visual stimuli differences in response time.

Context Effect

In addition to the four visual representation modes, a fifth mode was used as the baseline to determine how much people can make out based purely on context (an underlined blank in place of a visual mode of presentation). Results from ANOVA showed that all the four modes performed equally well and were significantly better than the baseline in which the verb is left blank (F(1, 94)=40.438, p < 0.001, η^2 = 0.301), Figure 5-20). This means that people actually tried to make out the concept from the representations instead of guessing purely according to the context. Furthermore, compared to the individual verb naming study, Figure 5-22 showed that sentence context had significant effect on improving the perception of the visualized verbs (F(1, 94) = 40.438, p < 0.001, η^2 = 0.301). Although the improvement gained from the context was not significantly different for all the modes (F(3, 282) = 1.242, p = 0.294, η^2 =0.013), videos still performed the best (F(3, 282) = 7.284, p < 0.001, η^2 =0.072). Detailed changes in comprehension accuracy introduced by providing sentence context were as follows. Other than “eat,” “play,” “read,” “stop,” “think,” and “write” which maintained high scores and “come,” “hope, ”remember,” “stay,” and “take” which were still low in accuracy, most verbs were significantly better interpreted. The five verbs with the biggest gain were “hear,” “find,” “thank,” “see,” and “wonder.”

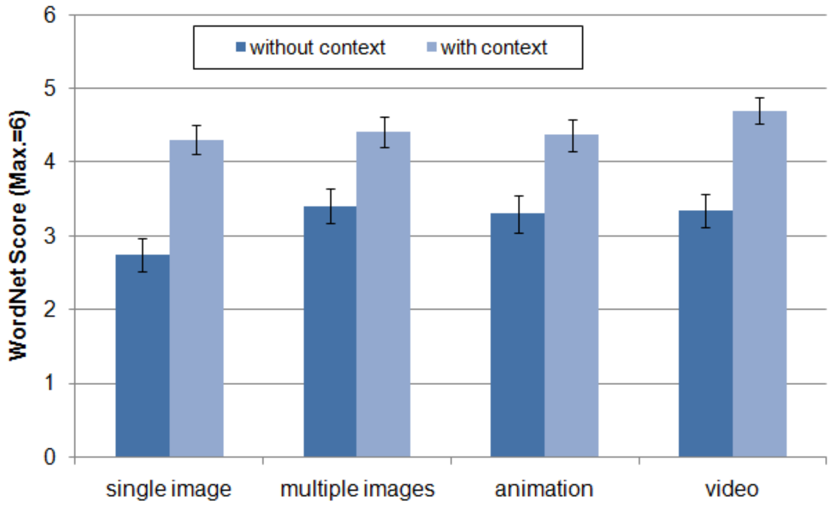


Figure 5-22. Influence of context on visual stimuli comprehension.

Verb Semantic Domains and Other Linguistic Properties

Although given the context, the evocativeness of the four visual stimuli across the nine verb domain was no longer significantly different (F(24,117)=0.779, p=0.730, η^2= 0.138), the video representations still significantly outperform the other stimuli on conveying contact and motion verbs (F(24,210)=2.165, p=0.02, η^2= 0.198). ANOVA results on various linguistic properties showed that abstract (F(2,380)=45.126, p<0.001, η^2= 0.106) and non-imageable (F(2,380)=102.135, p<0.001, η^2= 0.212) verbs were significantly harder to interpret. Particularly, the video representations were significantly better than the other visual stimuli on illustrating abstract and non-imageable verbs (Figure 5-23, 5-24). Polysemy (number of senses for each verb) did not show any significant impact on the comprehension accuracy.

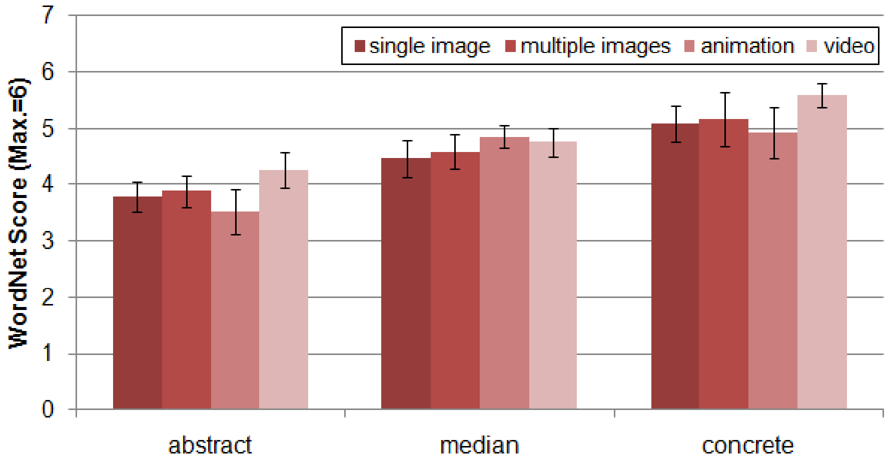


Figure 5-23. Comprehension accuracy differences by concreteness.

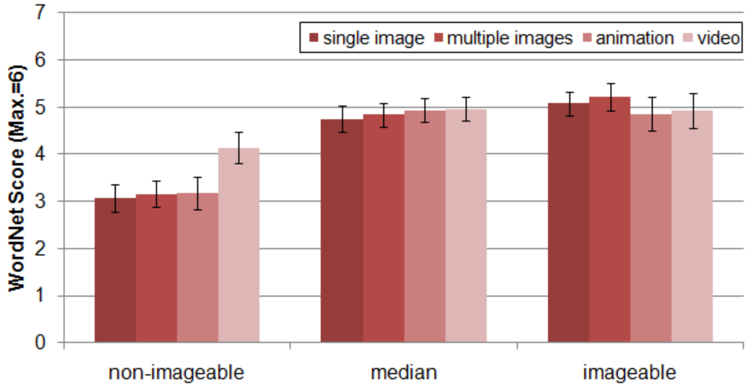


Figure 5-24. Comprehension accuracy differences by imageability.

Visual Cues

In the post-hoc test, again, the three visual cues (symbol, gesture, and facial expression) were examined. According to ANOVA results, symbol was the only factor that showed significant impact on the comprehension accuracy (F(2, 380)=32.356, p<0.001, η^2= 0.078). Although not significant, these visual cues affected the four visual stimuli differently. Compared to animations, videos benefited the least from direct symbols while received the least harm from indirect symbols (Figure 5-25). Videos made the best used of gestures (Figure 5-26) while were influenced more by facial expressions (Figure 5-27).

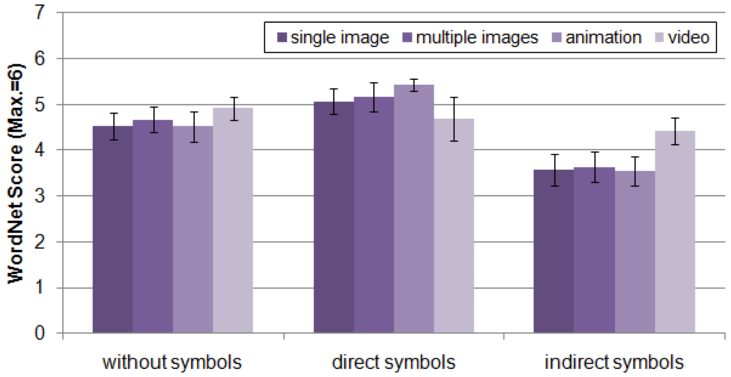


Figure 5-25. Comprehension accuracy differences with symbol.

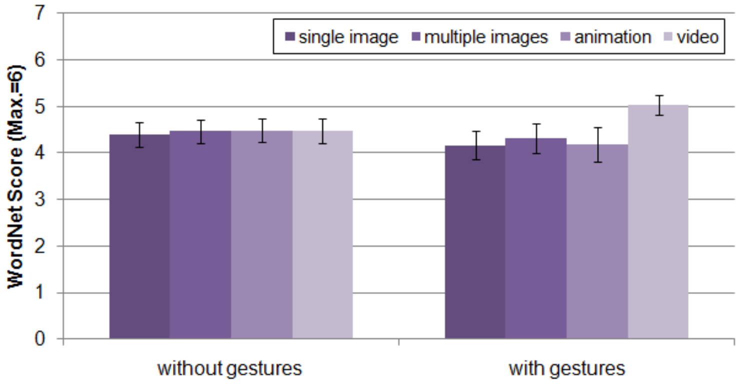


Figure 5-26. Comprehension accuracy differences with gesture.

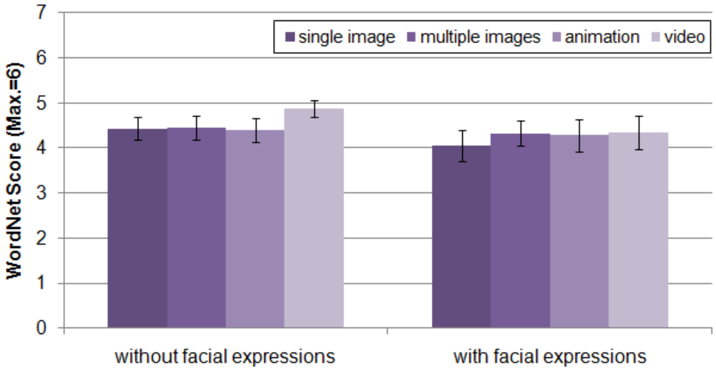


Figure 5-27. Comprehension accuracy differences with facial expression.

Sentence Comprehension Performance

Not only was the interpretation on individual verbs evaluated, but also the whole sentences. The score of the response to a sentence was the mean of the WordNet score of all the tested verbs in the sentence. Figure 5-28 (WordNet score) and Figure 5-29 (response time) show the detailed comparison of all visual modes as well as the baseline. Though outperforming the baseline, there was no significant difference among the modes in sentence interpretation (F(3, 258) = 2.054, p = 0.153, η^2 =0.008). No significant difference was found in sentence response time across the four mode either (F(3, 258) = 0.0283, p = 0.867, η^2 =0.0001). However, according to notes by the investigator and feedback from the participants, sometimes it was the images for nouns and adjectives that cost the most time. All in all, at the sentence level, video representations took slightly less time and were slightly easier to interpret.

The five most poorly interpreted sentences were: (1) Can I use your cell phone? (2) Let the lady get her coat. (3) I stayed out in the snow. (4) I come to say goodbye. (5) Keep the chocolate in the refrigerator. The interpretation accuracy for each sentence was consistent between young and old adults (correlation 0.8287). Since nouns and adjectives in the sentences were represented by single images, their perception (assessed by taking the mean of the WordNet score of all nouns and adjectives in each sentence) might impact that of verbs. However, there was no strong correlation (0.21) shown.

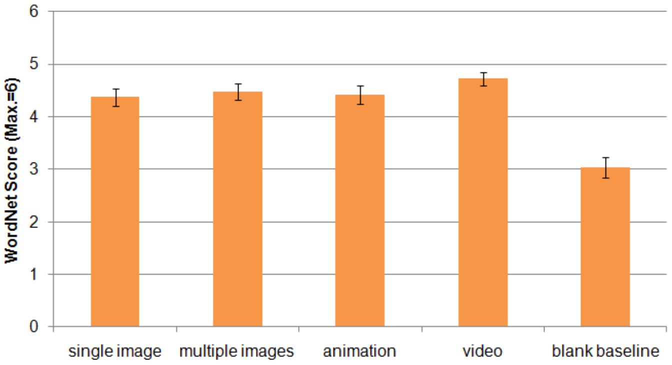


Figure 5-28. Sentence WordNet score.

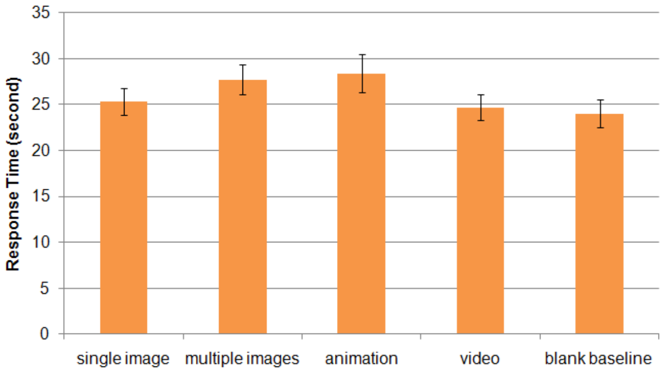


Figure 5-29. Sentence response time.

Subjective Feedback on Comprehension Difficulty, Speed, Confidence, and Preference

Based on the answers to the questionnaires, Table 5-13 summarizes participants' feedback on the four tested visual representations from various aspects, including easiness, response time (meaning how soon people start to have some thoughts, not necessary the final verb, popping up), confidence, and preference . Similar to the results from quantitative analysis, the difference was smaller than the feedback from the individual verb naming study across visual modes in terms of subjective feeling about easiness, response time, and confidence, and people's preference was not as strong.

stimuli	**preference**				**difficulty**		**speed**		**confidence**	
	1	**2**	**3**	**4**	**easiest**	**hardest**	**fastest**	**slowest**	**most**	**least**
single image	4	7	11	3	6	6	12	4	3	4
multiple images	6	5	3	11	8	9	3	6	9	6
animation	7	6	7	5	4	5	5	5	4	7
video	8	7	4	6	7	5	5	10	9	8

Table 5-13. Number of participants commenting on the difficulty, speed, and confidence of visual verb comprehension in context as well as personal preference.

5.2.3.5 Visual Verbs Study II (Cued Comprehension), Conclusions

In the comprehension of visual verbs in context study, the interpretation accuracy was greatly improved over individual verb naming (hypothesis H3 confirmed). The performance differences across the four stimuli were smoother, but the video stimulus still outperformed the single static image stimulus (H1 confirmed). There was a significant age difference in both comprehension accuracy and response time. For elderly participants, multiple images took the longest time to interpret and animations took the shortest time, whereas the response time for the four visual stimuli was not significantly different for the young participants (H2 partially confirmed). The interpretation accuracy difference given sentence context across various verb domains was different from without the context (H4 confirmed). The video representations were still the most evocative for contact and motion verbs as well as non-imageable verbs. Similar to the individual verb naming study, videos were less affected by symbols, benefited more from showing gestures, and were slightly harmed by revealing facial expressions (H5 confirmed). Interpretation accuracy of the other parts of speech in the sentences only had small correlation with that of the visual verbs (H6 partially confirmed). People showed slight preference on dynamic pictures over static pictures, but not as strong as in the individual verb naming study (H7 confirmed).

5.2.4 Discussion

There are other cases which can lead to failures in guessing the target verbs caused by people's perception of visual representations.

Misapprehension: People assigned concrete names to abstract symbols or mistook certain objects or expressions for something else. It happens in animations. For example, three out of eight people interpreted the cube in “give” as money, paper, or a cup of coffee, four out of eight people considered that in “take” a cup or a glass, and cards in “bet” were considered by two out of eight people to be glasses used in magic trick. Two out of eight participants mistook the worrying look in the video for a smile. One person mistook a public phone for a parking meter.

Distraction: Words in images and animations are distracting since people are always attempting to figure out the texts (for example, five out of eight people read the words in the “need” image, and five out of eight people spent extra time on decoding the random characters in the animations for “read” and “write”) no matter whether they have something to do with the main content or not. Sometimes the facial expression in images and videos for non-cognition and non-perception verbs are misleading too, for instance, four out of eight people came up with feeling related verbs for “come.”

Incompleteness: This is more common in video. People watched part of it and responded too quickly. For example, in the “pay” video, the person first counts the money and then gives it to someone else. Four out of eight people said “count” immediately, though three of them corrected their answer after they finished watching. The animation and video for “thank” is another example. Half of the eight people only saw that the person is being helped or receiving a gift and came up with “accept” or “receive.”

Wrong Focus: People sometimes focus on the wrong part or character. For example, “ask” is guessed as “teach” by two people because they thought from the lecturer perspective, and two people misinterpreted “give” as “receive”.

Familiarity: Several images for “stay” are someone training a dog to stay. Only one participant got it (immediately when she saw the picture) because she trains her dog like that. One participant was confused by the lock in the video for “remember” and guessed it to be a door bell since she had never seen one before.

Imagination: Some of the participants have vivid imaginations which might sometimes lead to unusual responses. For example, in the “give” image, keys are handed over. Two people imaged that these are car keys and came up with lend and exchange.

When participants encountered the situation that their interpretation of the verbs given visual representations did not fit in the sentence they made out, they tried to choose or balance between submitting the response regardless of the context and thinking of a new answer that makes the whole sentence makes sense, either of which could lead to a false response. In general, there are two cases that could cause the situation described above.

Confusion of the intended verbs: This happened when people failed to perceive the tested verb(s). For example, two people interpreted “Let the lady get her coat.” to “Direct the lady pick up her coat.” The response sentences did not make sense but they put down what they got from the representation. In another example, two people mistook “Keep the chocolate in the fridge.” to “Eat the chocolate in the fridge.” in which case they just submitted something that made the whole sentence meaningful.

Confusion with other parts of speech: Although there was no strong correlation between the accuracy of verb interpretation and interpretation of other parts of speech, mistakes in perceiving nouns and adjectives still causes failure in figuring out the verbs. For example, four people considered “say” in the sentence “I come to say goodbye.” to “sing,” and “announce,” “attend” because they mistook the “goodbye” image which shows the groom and bride waving goodbye to the crowds to “wedding.”

5.2.5 Conclusions

This chapter evaluated how well four different visual presentations (a single static image, a panel of four static images, an animation, and a video clip) can convey verbs in different domains that are most frequently used in daily oral communication with and without context. Whether the choice of visual mode affects the interpretation in the same way for young and old adults was also examined.

Results from the individual verb naming and the comprehension of visual verbs in context studies showed that visual representation modes perform significantly different, with videos the best among all. To be more specific, young adults did a similar job with multiple-picture representations (multiple static images, animation, and video) beating single static image, while for old adults video significantly outperformed the three other modes. Although there was an age-related decline in the accuracy and efficiency of verb retrieval, the influence pattern across different visual representations is similar for both young and old adults. It implies that when designing appropriate visual representations for verbs, the age of the target group should be taken into consideration. As for young adults, designers can choose multiple-picture representation with lower cost, while for old adults, in terms of effectiveness, video would be the best choice. Further analysis showed that for all four modes, the performance dropped as the imageability of the target verb decreased. Also, different modes are good at different domains of verbs; the most obvious is that videos did a significantly better job with contact and motion verbs which involve concrete actions and motions. This suggests that the choice of visual representations for verbs could be adjusted according to what kind of verbs are the most common in the intended application.

Exploration of various visual cues in four modes provides some guidance for the design of visual representations. For instance, designers should be careful with the use of symbols in animation and the facial expression on the actor/actress face in the video, which could be misunderstood. On the other hand, gestures helped people to interpret certain verbs.

Context significantly improved verb retrieval, though sometimes misinterpretation of other parts of speech might cause confusion. The general difference across verbs was no longer significant, but a similar pattern of performance was still found if verbs were categorized by their imageability. The negative impact of facial expression was washed out, but gestures and indirect symbols still have strong effects on interpretation.

5.2.6 Guidelines for Designing Visual Representations for Verbs

Based on the two studies comparing the efficacy of four visual stimuli for verbs, a list of guidelines for better visual verbs design was generated:

(1) Minimize distraction. Keep the background as clean as possible to help the viewers emphasize on the action or event on the foreground.

(2) Careful use of props. Keep the objects that are used to demonstrate the interaction as simple and generic as possible. Otherwise people are likely to associate the scene with a specific life experience (restrict imagination).

(3) Provide clear hint on the direction and intention of the action. Avoid people focusing on the wrong subject or event.

(4) Careful use of symbols. Universally accepted symbols can enhance people’s comprehension (Figure 5-30 left), however, symbols that are not directly linked to the intended meaning can cause confusion.

(5) Omit facial expression if not necessary. Keep a neutral face or not showing any detail of the face if the intended verb has nothing to do with emotions and feelings.

(6) Utilize common gestures/poses (not sign languages) (e.g. “think” in Figure 5-30 right).

(7) Careful use of special effects (images and videos). Not all of the people (especially elderly people) are familiar with special effects, and thus confusion may occur.

(8) Use familiar scenes and be sensitive to cultural differences (with symbol and gesture).

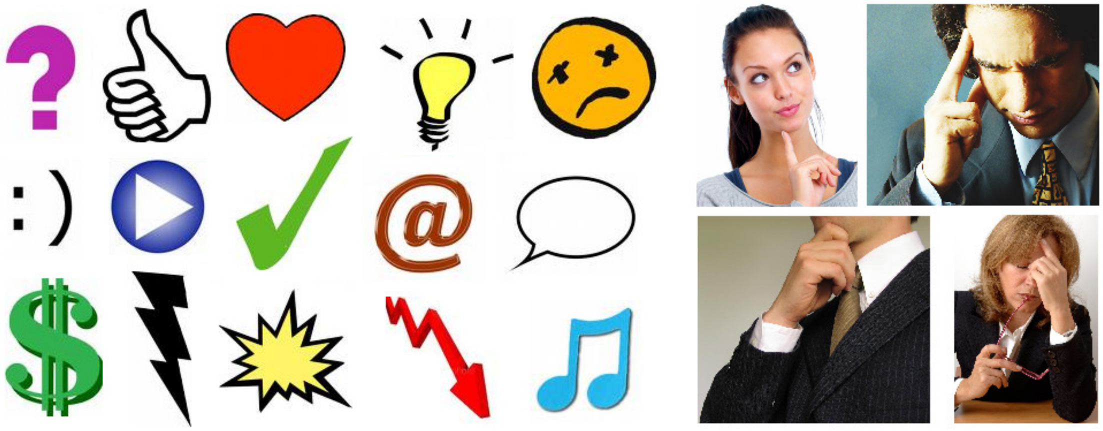


Figure 5-30. Examples of common symbols and gestures/pose (for “think”).

Chapter 4 and 5 discussed visual representations for conveying common concepts. However, not everything can be visualized. There are some concepts, such as “thunder,” can be better evoked through a natural sound. Therefore, the next chapter introduces a new kind of multimedia representations that is under-investigated: environmental sounds.

CHAPTER 6 NON-SPEECH AUDITORY REPRESENTATIONS FOR ALL PARTS OF SPEECH

Chapter 4 and Chapter 5 introduced the use of static visual representations (i.e. images and icons) and dynamic visual representations (i.e. animations and videos) respectively as supplements or extensions of words in communication. Studies have verified the efficacy of visual representations in illustrating a large variety of commonly used concepts. However, the experiments have also identified many occasions in which visual communication is not as effective, such as more abstract or less imageable concepts. This calls to question whether there are other stimuli that trigger different aspects of information perception from vision. To answer this question, I explored non-speech environmental sounds were explored as a potential candidate for carrying concepts for communication.

One disadvantage that auditory representations have over visual stimuli is that sound requires time to play and has to be played in sequence (Yost2007). Many concepts do not produce a sound, or a distinctive sound. However, there are still cases where a sound can effectively evoke a concept even better than a picture. For example, “thunder (unlike lightning)” and “chirp (unlike bird)” are harder to visualize; “coughing” and “sneezing” can be distinguished more easily by their sounds than by pictures; and events like “tuning a radio” can be better portrayed via a sound unfolding over time than a static picture. Therefore, thorough research needs to be conducted to understand the communicative power and limitations of environmental auditory representations.

6.1 Auditory Perception and Sonification

Before any discussion, definitions to some related concepts used throughout the chapter are given in this section. First of all, **sonification**, as appearing in the title of this section, is defined as the use of non-speech sound to convey information (Kramer et al. 1999).

Sound is a “travelling wave which is an oscillation of pressure transmitted through a solid, liquid, or gas, composed of frequencies within the range of hearing and of a level sufficiently strong to be heard, or the sensation stimulated in organs of hearing by such vibrations” (dictionary2006). A sound is the auditory effect generated by a given cause. Certain kinds of sounds can be detected by the sensation of hearing. As the subject of my research, the word “sound” appearing anywhere in this thesis refers to this sub-category of sounds that people can hear. **Audio**, in the context of this thesis research, refers to audible sounds (also called acoustic signals). Sounds can be roughly divided into three categories: speech, music, and environmental sounds. I consider singing as a combination of speech and music.

Speech (spoken words) is the vocalized form of language and communication. A speech sound is "(phonetics) an individual sound unit of speech without concern as to whether or not it is a phoneme of some language" (by WordNet). **Music**, by definition in WordNet is "an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner," though in some context "music" is used as a generic name for any pleasing and harmonious sounds. Speech and music (in its narrower sense) are means to communicate by their nature. Besides spoken languages, people are rather familiar with expressing attitudes and emotions through music as well. A great piece of music, like good literature, is rich in information, thoughts, and feelings. People not only listen to music for entertainment, but also use it to share certain messages.

Environmental sounds (in a coarse definition) include all the non-speech non-music auditory signals that people encounter in daily life. Examples of environmental sounds are people laughing, foot steps, knocking on the door, bottle clinking, bird song, wind blowing, engine roaring, ambulance siren, telephone ringing, etc. Unlike speech and music, in most of the cases environmental sounds occur naturally without the purpose of information delivering. Environmental sounds are part of everyday experience. There are indeed situations where the environmental sound does mean something, which will be discussed in the next section. Other terminologies have been used to denote environmental sounds in this document, including non-speech sounds/audio, nonlinguistic sounds/audio, everyday sounds, and natural sounds (though it sometimes refers to sounds related to animals and natural phenomena only in other materials). Environmental sounds, not speech or music, are the subject of the research presented in this chapter.

6.1.1 Auditory Communication via Environmental Sounds in Everyday Life

On one hand, there are nonlinguistic sounds that are made deliberately to convey specific information in everyday life. For example, car horns, phone rings, and doorbells are used to catch people's attention. Fire alarms and foghorns can alert danger. The "ding" and "buzz" sounds are widely used in TV programs as notification of right or wrong. Sounds have been used to signify the start or end of certain events as well (e.g. elevator arriving at a floor). As a side note, audio-encoded languages like the Moss Code do not fall into the scope of discussion in my thesis research. Auditory communication as exemplified above has several advantages. First, audio signals (especially with distinguishing pitch and/or high volume) can easily attract people's attention. Humans have evolved to be very sensitive to acoustic signals that stand out from the surroundings. Second, people can receive the information without shifting their focus, which means, eye-free. For example, drivers can learn that some car is passing from behind without turning around. Third, compared to visual communication, sounds have fewer constraints in terms of distance in effect and requirement of obstacle free. Unlike light, sound waves can travel through and around opaque materials. A foghorn is a good example in this case. Fourth, audio can affect multiple people loosely distributed at the same time, which is also an advantage over visual signals. For instance, people on the same floor or even in the same building can hear the same fire alarm. These kinds of sounds can be considered as auditory symbols/signs that are used in convention, though some learning may be required.

On the other hand, besides audio signals that are designed for information delivery, people use sounds to imply other commonly known messages as well. For instance, people sometimes fake a coughing sound to signify that someone is uncomfortable or ill. As another example, in comedy shows we often hear audience laughing in the background indicating that it is supposed to be a funny scene. Sounds can be effective in these cases because through daily experience, people have already built up certain associations between the unique sounds and events (usually no deliberate learning is involved). Hearing sound can trigger the recall of the connected experience. The associations may come from natural process (e.g. bottle broken and the glass shattering sound, being hungry and stomach grumbling) or frequently co-occurring events (e.g. Christmas and jingle bells). In the next sections, I will introduce the theories behind the human perception of sounds (especially environmental sounds) through hearing.

6.1.2 Everyday Listening and Musical Listening

Each sound contains two sets of information, one about its origin, and the other about its acoustic properties. Information of sound origin refers to the event that produces the sound (vocal cord vibration for speech sound; striking surface or plucking strings for music; and other contact or friction events for environmental sounds). Acoustic properties of a sound refer to characteristics like pitch, volume, and temporal pattern. Depending on which information the listener is focusing on, listening can be divided into **everyday listening** and **musical listening** (Gaver1993).

In musical listening, people capture the pitch, loudness, timbre, and changes of the sounds (Figure 6-1). The listeners are intending to identify and interpret these low-level auditory attributes. Musical listening, as suggested by its name, occurs more often in the creation and perception of music. People discriminate different music pieces by their time/frequency information.

On the contrary, everyday listening (Figure 6-2) is the perception of auditory events (e.g. the characteristics of the sources of the sounds, their position and interactions involved). Research (Handel1989)(Jenkins1985)(Vanderveer1979a)(Vanderveer1979b) has shown that people can identify significant aspects of environmental sounds from their experience. For instance, people can tell a car engine sound from footsteps on a wooden floor, and detect if the car is approaching or departing. The details on how people perceive this information is discussed in the next section.

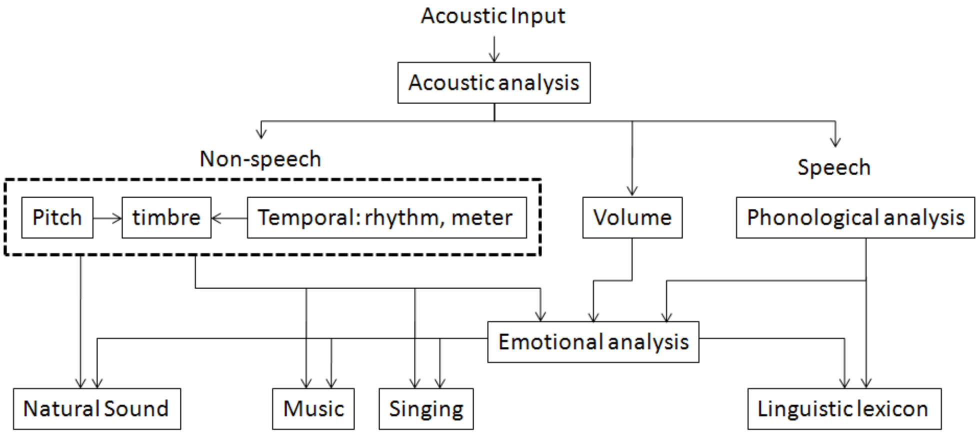


Figure 6-1. Musical listening.

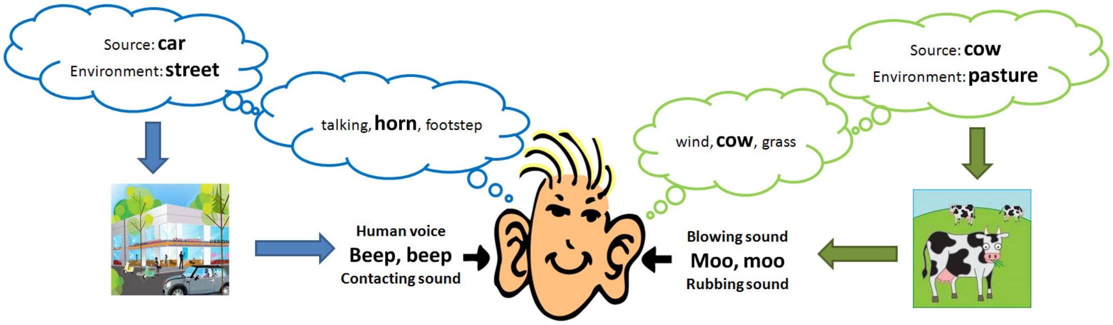


Figure 6-2. Examples of everyday listening.

People can switch between musical listening and everyday listening with the same sound depending on the context, expectation, and prior knowledge. For example, one may say “I like the song because of the melody and the rhythm (musical listening), plus, the guitar solo is awesome (everyday listening).” In some other occasions, people may start with everyday listening, trying to identify the source of the sound. If cannot tell, they may switch to more abstract description or even using acoustic features as a reference (e.g. “Did you hear the explosion?” vs. “Did you hear that deep loud noise?”). Research (Houix et al. 2007 part2) has also shown that different people may be better at different kinds of listening. For instance, music specialists describe and categorize everyday sounds based more on musical listening, whereas ordinary people use everyday listening. Although both everyday listening and musical listening can take place in perception of speech, music, and environmental audio, for the purpose of research, discussions will be limited to environmental sounds from this point on.

The use of environmental sounds to convey everyday concepts was explored in the work presented in this chapter, for the purpose of facilitating language comprehension and communication across language disabilities and barriers. The intended concepts are often directly associated with sources, environments, and interactions involved in the creation of the sound, and thus can be perceived and evoked through everyday listening. There are a few places musical listening can come into play, e.g. depicting "distance." Research (Clarke1996)(Dick2002) has shown that many people with language disorders still maintain the ability to identify natural sounds. This suggests that everyday/musical listening is viable for both healthy populations with limited language skills and language-impaired populations. Environmental auditory representations have the potential of assisting language comprehension and expression.

6.1.3 Perception of Sound Events and Related Properties

In everyday listening, once an acoustic signal is received by hearing, listeners start the recognition process from the top down or from the bottom up or both (Ballas1987). A top-down process begins with expectations and prior knowledge that suggest possible candidates for the sound production events, and then searches in the memory space based on analysis of the input to narrow down the options. For example, if a person sees lightning, he/she knows that by default, the soon-to-come loud noise is thunder. On the contrary, a bottom-up process reconstructs the auditory scene from every piece of information retrieved from the input sound and its context. For instance, the listener hears a sudden loud noise with echoes. It sounds like the air has been torn apart. He/she constantly hears water dripping sound outside the window as well (the context). Then he/she may conclude that the loud noise was thunder. In the scenario of Augmentative and Alternative Communication with nonlinguistic auditory representations, it is less likely that people know in advance what sound to expect, and thus, the following discussion focuses more on how people perceive sounds in a bottom up manner.

The events happening around us that produce a sound are called sound events. After the sound reaches their ears, listeners form a cognitive representation of the sound event, and this representation is called an auditory event (Houix2007a). As in Figure 6-2, car honking is the sound event, the horn sound is what is produced, and perception of "car horn" in the listener's head is the auditory event. Each environmental sound has certain acoustics characteristics that can help with recognizing the sound event, similar to the timbre of music. Some of the characteristics are more recognizable than the others, they are often the clues that listeners look for in order to generate a reasonable auditory event. The characteristics can be related to the physical property of the sound-producing object (e.g. size, shape, and materials), can be related to frequency, can be related to temporal pattern (e.g. dot matrix printer), can be related to sequence of simple sounds (e.g. the sound sequence of key inserting and turning, to turning-over, to engine running suggests the event of starting a car), and more often, can be a combination of all the factors listed above. This is called the multidimensional nature of an environmental sound, and has been studied for several kinds of everyday sounds, such as natural phenomena (Bjork1985), air conditioning noise (Susini et al.2004) and car horn (Lemaitre et al.2009).

An auditory event usually contains information more than just the identity of the source of the sound. Properties of the source, interactions involved, surrounding environment, and even how the sound can be categorized and related to other concepts can be components of an auditory event (VanderveerThesis1979). Studies have shown that people can tell from a sound the material (Lakatos et al.1997), the shape (Kunkler-Peck and Turvey2000), the size (Carello et al.1998), and the hardness (Freed1990) of the sound source object(s), and even the gender of the walker (Li et al.1991). Through indentifying individual sound events co-occurring in a scene, people may be able to recover the entire picture of sound source(s) and the surrounding environment as well (Figure 6-2). Research that has been conducted on environmental sound perception mainly falls into two categories: identification and classification.

6.1.4 Identification of Environmental Sounds

Systematic research has studied how well/quickly people can identify an environmental sound and the factors that may influence people's perception or decision.

6.1.4.1 Confrontational Naming

Confrontational naming is a common method used in experiments of sound identification (Vanderveer1979)(Ballas1993) (Houix2007b). Usually a set of (dozens to about a hundred) short sound clips (less than a second) were played one by one to the participants. People were asked to name a noun or a verb denoting the cause of the sound. Descriptive narration was forbidden. In some of the studies, the participants pressed a key to start the sound and pressed it again to indicate the availability of a conclusion, giving a measure of response time. This kind of study setting provides some insight into the bottom-up process of auditory perception. Combined with other work that looked into the top-down perception of everyday sounds (Howard and Ballas1980), it was concluded that human listeners utilize both perceptual information derived from the sound and prior knowledge to identify an everyday sound.

6.1.4.2 Confusion Errors and Factors of Impact

However, not every sound is distinctively recognizable. Homonymous sounds refer to a group of sounds generated by different events but share the same acoustic properties. It is similar to the homonymous words situation in language, i.e. different words sharing the same pronunciation, such as “sun” and “son.” Engine sound from different vehicles (an airplane, a car, or a motorboat) is an example of homonymous sounds. Studies (Ballas and Mullins 1991) showed that context can help differentiate homonymous sound. In the engine sound example, wind vs. road vs. water gives a good clue of which vehicle it is.

There are other occasions where the sounds are not exactly alike, but are similar enough to cause confusion. It could be sounds that involve the similar materials (e.g. iron and steel), sounds that involve similar interaction (e.g. friction in striking a lighter and cutting with scissors), or sounds appearing in similar context (e.g. lock and unlock). Listeners often generate several guesses and decide accordingly. If they choose the wrong option, a confusion error occurs.

It has been suggested that the fewer options a sound can introduce, usually the faster and more accurately one can identify the sound (Ballas and Sliwinsky1986). The listeners' familiarity of the sound scene and the ease of mental image generation also have big influence on their sound naming performance (Ballas1993).

Previous studies on identification of environmental sounds focused on perception of the sound events. They have not yet touched the area of conveying information through environmental sounds. The audio stimuli used in the experiments were too short to evoke extensive message. The sounds presented in the naming studies were only a subset of the non-speech audio examined in this research. Other factors (e.g. parts of speech) that may affect people's interpretation of environmental sounds are also explored.

6.1.5 Classification of Environmental Sounds

Classification of Environmental Sounds is studied through two methods. The first method looks at the sounds themselves and tries to understand the nature of the sound producing events. The second one approaches the problem via experiments, by examining how people categorize the sounds they hear.

There are two kinds of models developed via the first method. A top-down taxonomy proposed by Gaver (Gaver1993) first decomposes specific sound events by the materials involved in the sound production, i.e. vibrating solids, gases, and liquids, and then, the simple interactions that are always associated with each material (e.g. impacts for solids, explosions for gases, and splashing for liquids). This taxonomy is rather natural, since people are less likely to confuse solid sounds with water sounds or air sounds, although the simplicity limits its expressiveness. The Closing the Loop of Sound Evaluation and Design (CLOSED) Project (Houix2007a) constructs a bottom-up structure for environmental sounds based on Gaver's taxonomy. At the bottom level, sound models that describe the physical cause of the sounds, such as an *impact model* and a *friction model*, are developed. One level up, textures and associated basic events are assigned to each sound model. For example, rolling on solid surface is assigned to the *impact model*, whereas rubbing between two solid surfaces is assigned to the *friction model*. The interaction of these basic events yields the next level with more complex sound generating process. For example, sliding is a process derived from the interaction between rolling and rubbing.

A lot of empirical studies have been conducted to study human categorization of environmental sounds, such as (Guyot 1996)(Scavone et al. 2001)(Ballas1993)(Houix2007b). The basic idea is to have people sort or group a set of environmental sounds and then explain the strategy used in the classification. It may suggest how everyday sound events are connected in human perception space. There are two common strategies people used for constructing the sound categories. One is by the similarity in acoustic characteristics perceived through musical listening, such as pitch (e.g. deep horn sound and low-pitch explosion sound), temporal changes, and mechanical vs. electronic. The other is by the source(s) and movement(s) involved in the sound production as a result from everyday listening.

In my study of how well environmental sounds can evoke everyday concepts, different sound classification schemes (based on source, interaction, and environment) are used to examine if there is any relation between the nature of the sound categories and their evocativeness (details see *Section 6.3*).

6.1.6 Sonification

Research in everyday sound identification and classification has shown that people have the ability to perceive sound events, and thus, it is viable to use audio signals to convey certain information. Sonification refers to the use of acoustic signals to illustrate data and information. Compared to visualization, non-speech audio has been found to have the advantages of evoking temporal characteristics and illustrating transformation over time (Hartmann1997) (Moore1995)(Moore1997). Furthermore, auditory display does not require users to direct their visual attention, and thus is suitable for eyes-free environments. Sonification applications mainly appear in two fields: audio systems in industry and auditory interfaces for computers.

6.1.6.1 Audio Systems in Industry

Sonification techniques have been applied to catching attention, alerting to danger, and depicting changes in data by the shift of sound frequencies and intensity (Figure 6-3). For instance, Begault et al. (Begault1996) and Patterson et al. (Patterson1980) developed audio alert and guidance systems for airplanes, automotive, factory, etc.). The multi-level alarming system in the control room of a nuclear power plant (Visuri) is another example of such auditory systems. Gaver et al. (Gaver1991) revealed the importance of auditory feedback in human-system interaction for diagnosing problems and monitoring system functions in a plant. Scientific data analysis (Pereverzev1997) is another application of sonification. In these auditory systems, only simple synthetic audio patterns are used. The information contained in the auditory representations is limited.

Figure 6-3. Examples of industrial auditory system: power plant control room (left), airplane navigation system, and alarm system.

6.1.6.2 Auditory Display for Computer Interfaces

Another area that applies sonification is illustrating computer interfaces via audio patterns, e.g. earcons and auditory icons. Earcons (Figure 6-4 right) are non-speech synthetic audio patterns designed to provide information about objects, operations, status, and interactions on computer interfaces via auditory features like pitch, rhythm and volume (Blattner1989)(Brewster1998). People are not familiar with synthetic sounds and their assigned meanings, and thus the use of earcons requires learning.

Compared to earcons, auditory icons are more natural since they encode computer events with everyday sounds. SonicFinder (Gaver1989) was a computer interface exploring the use of auditory icons. It mapped sounds like bouncing and breaking to computer events analog to everyday events (Figure 6-4 left). Many people (Mynatt1994)(Garzonis2009) have continued the work with auditory icons ever since. Auditory icons are more intuitive and accessible than earcons (Garzonis2009), although it still restricts the information conveyed to computer related experience.

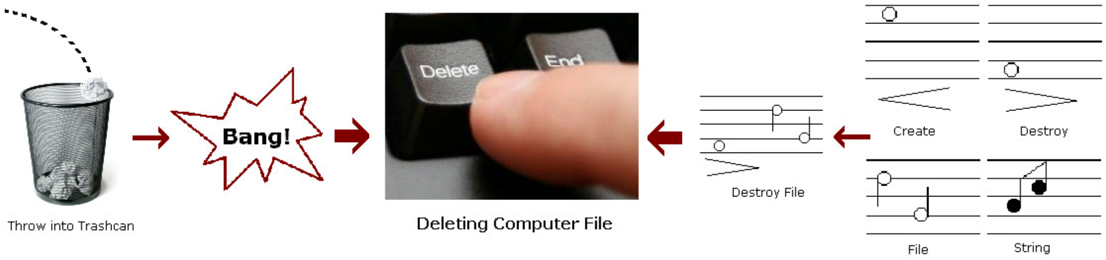


Figure 6-4. Auditory Icons (left) and Earcons (right) for deleting computer files.

Both earcons and auditory icons aim to represent specific information, mainly concentrating on computer interfaces and activities. Earcons and auditory icons are metaphors or analogies, instead of a direct translation of the everyday experience embedded in the sounds. My thesis research differs in that the auditory vocabulary is extended to a scope that can reflect topics in daily life, and in that the potential users are populations with communication difficulties.

6.2 SoundNet Overview

Since sounds and pictures are two different information channels, and their evocativeness varies for different concepts, we keep the visual component and auditory component of the multimedia AAC vocabulary separated. Another reason to build the sound augmented vocabulary independent from the pictorial part is, the sound and picture selected for a given concept based on prototype theory or effectiveness may not match each other. For example, the picture of a parrot may be picked for the concept "bird," while the chirping sound of a linnet could have been selected. People may get confused when they see the parrot picture while hearing a linnet singing in the background.

SoundNet (Figure 6-5), the auditory vocabulary being built separately, is a semantic network with environmental auditory representations associated with concepts. Different from auditory icons and earcons, SoundNet consists of concepts commonly used for everyday communication, not just restricted to computer events and elements. SoundNet utilizes people's everyday listening ability to evoke information related to the source and environment of the sound production. No second-level mapping as in auditory icons is involved. The construction of SoundNet was divided into several steps: first, identifying a potential inventory of frequently used concepts that may be illustrated by a natural sound; second, determining what sound scene may evoke the concepts in the inventory; third, create the auditory representations accordingly.

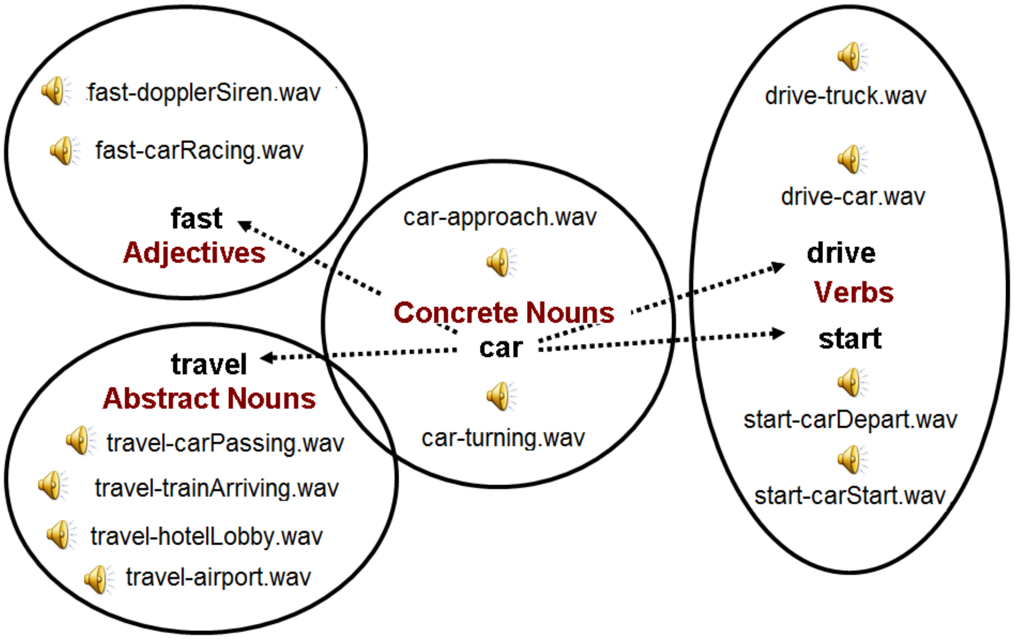


Figure 6-5. Examples of SoundNet.

6.2.1 Concept Inventory Construction

The goal of building SoundNet is to use non-speech audio to improve language comprehension and acquisition for people facing language disabilities or language barriers. In a broader scope, SoundNet can not only facilitate daily communication, but also help with language learning and language rehabilitation. The “core vocabulary” designed for SoundNet, therefore, should include words needed to discuss common topics in daily communication and should cover the major parts of speech. The initial selection of words came from the Lingraphica glossary (lingraphica), since it is designed to support communication for people with aphasia and has been put in practice for over 10 years. Lingraphica currently has over 3,600 words and phrases. After eliminating duplicates and symbols (e.g. characters and punctuation) and stemming, 1376 words were retrieved.

However, there is no guarantee that all of these words from Lingraphica can be represented by a sound. To further refine the selection, the initial list was compared to the collection of words generated from the BBC Sound Effects Library (BBC) captions. BBC Sound Effects Library is a large environmental sound dataset, and it is introduced in more details in the S*oundnail Creation* section. The idea is, if a word appears in one or more sound captions, it is likely to be relevant to the auditory scene. BBC sound captions are in the form of short sentences describing the sources and events involved in the sound production. The captions were divided into bags of words, from which the functional words were removed and the content words were stemmed. This yielded 1368 words. The overlap between the Lingraphica and BBC glossaries generated the core SoundNet vocabulary. The senses of the words in the compressed list were determined based on the assigned icons in Lingraphica. For those words which can have multiple parts of speech (like “walk” and “water”), only the more frequent sense based on WordNet (WN) was kept. The final concept inventory included 211 nouns, 68 verbs, 27 adjectives, and 16 adverbs.

6.2.2 Audioability Rating

Audioability, a word coined by us, refers to the ability for a concept to be conveyed by an environmental sound. Audioability is a measurement of the tightness of the association between concepts and their potential auditory representations. Intuitively, almost half of the 322 concepts in the SoundNet inventory seemed to be difficult to illustrate via a sound. Five people from the Princeton Aphasia Project and the Princeton SoundLab assigned an audioability rating to each concept in SoundNet based on a four-point rating scale exemplified in Table 6-1. Two additional SoundLab members served as judges to help finalize the audioability scores. The concepts (184 out of 322) that received a score greater than 1 are considered **audioable**. The audioability rating helped to evaluate the candidacy of the proposed auditory representations for each given concept. If a concept is considered highly audioable but people cannot recognize it from the proposed sound, it means the sound is not suitable as a representation. On the contrary, if a concept is evaluated as “hard to convey,” the failure in the proposed the sounds might merely due to the linguistic property of the concept, and we may conclude that in general, it is difficult to evoke this concept through any sound.

Rating	Justification	Example
0	cannot make sound or be used to produce sound and cannot be represented by sound	“am”
1	can make sound or be used to produce sound, but cannot be represented by sound	“desk”
2	can make sound or be used to produce sound, and may be able to be represented by sound, meaning the sound could be ambiguous	“bear” and “glass”
3	can make sound or be used to produce sound, and can be represented by sound, meaning the sound is distinctive	“dog”

Table 6-1. Audioability four-point rating scale.

For each of the 184 audioable concepts, our graders and judges separately wrote a script of sound scenes that can be used to depict the concept. The scripts were refined through discussion, and they formed the guidelines for selecting and assigning sound clips based on their original captions. For example, for the concept “call (verb): making a connection by telephone,” people created scripts like “rotary dial phone,” “push button phone with touch tones,” and “phone ringing and being picked up.”

6.2.3 Soundnail Creation

The three sources of the environmental audio clips employed in SoundNet include the BBC Sound Effect library (about 2/3 of the representations), Freesound (freesound) and FindSounds (findsounds). Based on the scripts composed by our group members, over 300 audio clips were collected for the 184 audioable concepts. For three practical reasons, we constructed 5-second auditory illustrations called **soundnails** from the original sound files instead of directly utilizing the raw clips:

(1) The majority of the BBC sound effects are dozens of seconds long, and many even last several minutes; this is also the case for the clips obtained from Freesound and FindSounds. Long sound clips require extensive listening and processing time.

(2) These long clips carry richer and more complex information than can be conveyed by a single concept. For the applications we have in mind, such scripts are not suitable.

(3) The third problem is the size of the sounds. The audio clips, especially those from the BBC library, are high-resolution stereo files, which makes them difficult to load and play on webpages instantly, and also impractical to store on mobile devices.

For practicality and quality control, all selected clips were down-sampled to 16kHz, 16 bit mono, which is a sample rate at which people can still well recognize the sound scene. The 16kHz sample rate was determined based on the fact that many games (especially mobile/handheld) use 11.025 or 22.05kHz sample rate, and the speech recognition community has conventionally used 16kHz for recognizers. It was critical to keep the file sizes small for web transmission in the studies described in the sections below. Also, there is no guarantee that our target users and study participants would have the proper mpeg/other audio decompressor installed and working on their computers. Our group concluded that 16kHz, 16 bit audio was of acceptable quality, and this was verified in a pilot study (Scavone et al. 2001). All sound clips were randomly chunked into five-second fragments, as it was desired to keep files short, of the same length to balance the experimental conditions, but long enough to still embed enough information.

Signal Processing and Machine Learning algorithms were applied to automatically select the representative fragment (one that contains the most distinctive information). All five-second fragments were represented in the frequency domain, and six features (Mean and standard deviations of RMS Energy, Spectral Centroid, Spectral Flux, 50% and 80% rolloff, and 6-10 MFCCs) reflecting different characteristics were extracted (Table 6-2, Tzanetakis and Cook 2002). Later, for each original sound clip, all the fragments were clustered into three to four groups by the K-Means algorithm based on the extracted features (Figure 6-6). The fragment that was the closest to the center of its cluster was picked out as the representative for the group. Two methods for clustering and candidate selecting were experimented. In the first method, all the audio features for each sound fragment were combined into a single vector that is used in both clustering and representative selection (Figure 6-7 (a)). In the other method, the audio features were split into two groups. The energy-related features (RMS Mean and Standard Deviation) were used for clustering, while the other spectrum-related features were only used for determine the distance of fragments within each cluster (Figure 6-7 (b)). The split feature vector scheme can handle the sounds with clear energy change (such as the “Coffee Grinder Filled, Operated.wav” sound in Figure 6-6). In contrast, the combined feature vector scheme works better with sounds that have relatively uniform energy distribution but distinctive frequency distribution.

Feature	Description
Mean and standard deviations of RMS Energy	The average frequency that will vary for each signal.
Spectral Centroid	The average frequency of the signal weighted by magnitude.
Spectral Flux	How much the frequency varies over time.
50% and 80% Spectral Rolloff	How much of the frequencies are concentrated below a given threshold (50% and 80%).
6-10 MFCCs	Mel-Frequency Cepstral Coefficients: amplitudes of spectra specified by a set of filters.

Table 6-2. Audio features used in soundnail creation.

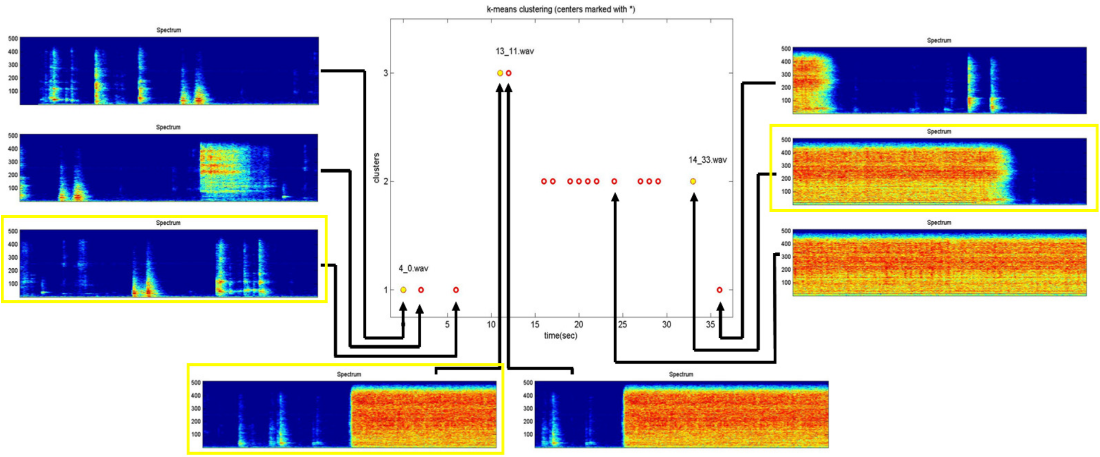


Figure 6-6. Sound fragment clustering for the sound “Coffee Grinder Filled, Operated.wav.”

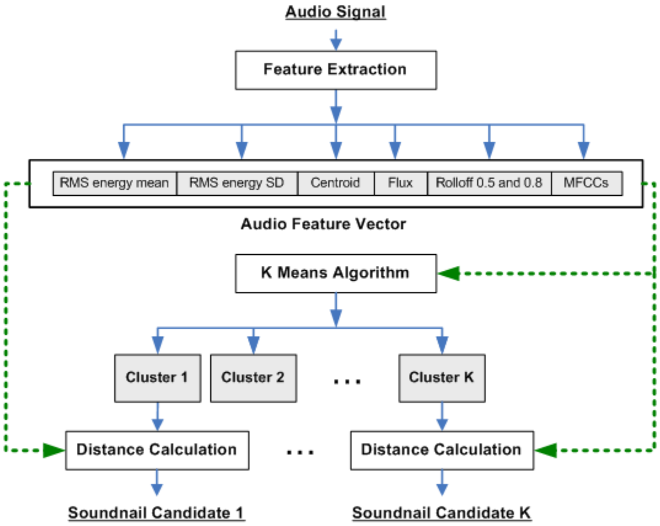


Figure 6-7.(a) Soundnail selection algorithm flowchart in combined feature vector scheme.

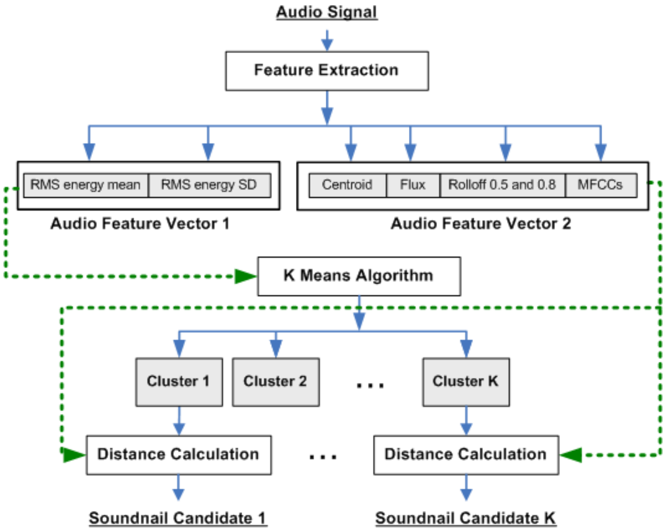


Figure 6-7 (b). Soundnail selection algorithm flowchart in split feature vector scheme.

After automatic processing, up to four soundnail candidates (depending on the complexity and variance of the sound scene) were generated for each sound clip. Each soundnail candidate included certain characteristic parts of the sound scene. In the last step, Aphasia Project and SoundLab members examined all candidates and selected one as the final representation to be associated with the intended concept in SoundNet. The decision was made according to the scripts. For example, 5-second fragments from the sound “BT Lines AND Tones, 3 STD Rings, Phone Answered With Pip.wav” were clustered into “connecting,” “ringing,” and “ringing and picked up.” The representative from “ringing and picked up” was assigned to the concept “call: get or try to get into communication (with someone) by telephone.”

A total of 327 soundnails were generated for the 184 audioable concepts in SoundNet. It is not a one to one mapping (Figure 6-1). Certain concepts are assigned to more than one sound, either from the same sound event but with different properties (e.g. for the noun “car,” the sounds “car approaching” and “car turning” are both from the driving event) or from completely different scenes (e.g. sounds from “airport,” “train station,” and “hotel lobby” may all depict the concept “travel”). On the other hand, some soundnails are used to depict multiple concepts. For instance, the soundnail “vacuum cleaner turned on” is assigned to both “vacuum (noun)” and “clean (verb).” Most of the soundnails were normalized in volume, except for those that explicitly needed to have higher or lower volume, such as the soundnail for “distance”.

6.3 Evaluation of Non-speech Auditory Representations, a Labeling Study on Soundnail Precision

6.3.1 Research Questions

Before adding SoundNet to the multimedia AAC vocabulary we are building, we need to investigate if the soundnails effectively convey their assigned concepts or cause confusions, and further, come up with guidelines to generate more evocative auditory representations. This can be extended to more general research questions: (1) what kinds of concepts can be evoked by a natural sound? (2) What kinds of sounds are distinctive enough to evoke a concept? (3) What kinds of miscomprehension may appear in everyday listening? And (4) what introduces the confusion? To address these questions, a study was designed and conducted to evaluate the precision of people's perception of these auditory representations. This was done through collecting human-generated semantic labels for the nonspeech soundnails in SoundNet.

6.3.2 Study Design

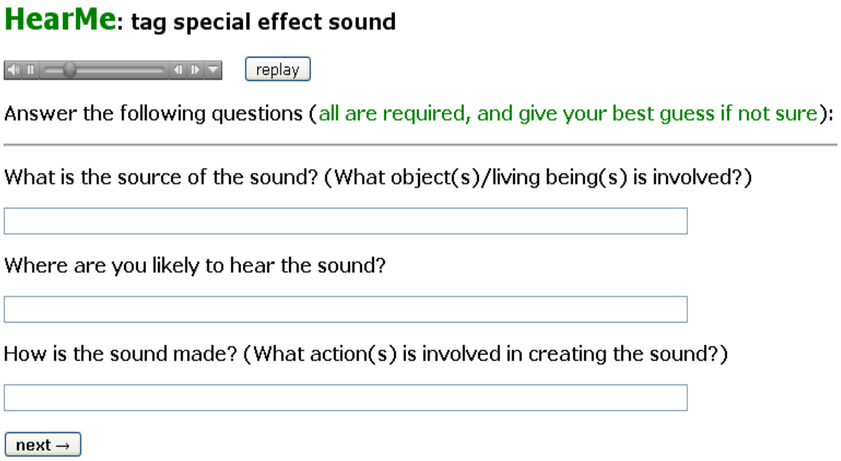
The goal of the study was not just to gather sound labels but to determine whether, and in which cases, specific aspects of the soundnails evoked responses. Thus, instead of acquiring a single label, we collected answers to three targeted questions, since people tended to label a sound with its source(s) in a free tagging study (Vanderveer1979)(Gaver1988). We hypothesize that in some cases, the location, the source, or the manner of the sound production is salient, but perhaps not all of these. We also wanted to see in which cases not all of the words in the label were named by the subjects. For example for the "walking on snow" soundnail, "walking" and "footstep" were generated, but not "snow," suggesting that the location was not audioable (can be evoked by an everyday sound) here. In the study, three questions targeting different aspects of each soundnail were asked (Figure 6-8), so as to encourage people to come up with as much information across different parts of speech as possible:

(1) What is the source of the sound? (What object(s)/living being(s) is/are involved?)

(2) Where are you likely to hear the sound?

(3) How is the sound made? (What action(s) is/are involved in creating the sound?)

The study interface was web-based. A pilot study was carried out with twenty-two Princeton undergraduate students from the Human Computer Interface Technology class to help modify the design. Five to eight labels were produced for each soundnail, and the time to label each soundnail was automatically logged as well. A post-study questionnaire was given to gather feedback on the design and interface of the experiment. Several aspects of the study interface (e.g. auto-play of the sound and phrasing of the questions) were modified according to the feedback gathered in the pilot study. It also generated the ground truth human labels for quality control of the actual large-scale online study.


HearMe: tag special effect sound

Answer the following questions (all are required, and give your best guess if not sure):

What is the source of the sound? (What object(s)/living being(s) is involved?)

Where are you likely to hear the sound?

How is the sound made? (What action(s) is involved in creating the sound?)

Figure 6-8. Soundnail labeling interface.

6.3.3 Study Setting and Conduction on Amazon Mechanical Turk

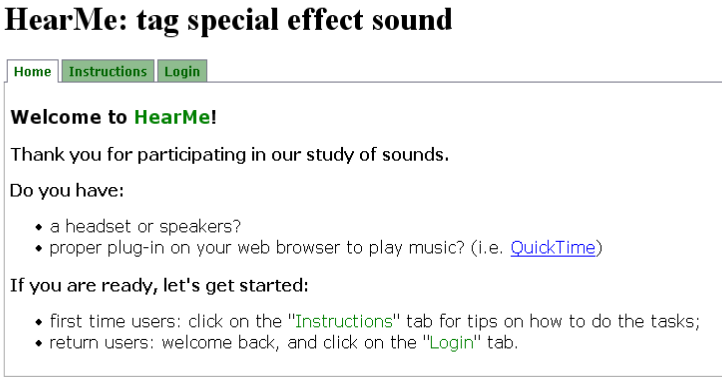
For the purpose of gathering a large amount of human data, the actual study utilized the service and environment provide by Amazon Mechanical Turk (AMT). AMT is a web platform operated by Amazon, where people can post web-based surveys in which people all over the world can take part, requiring only an Amazon account. This provides a large human subject pool. AMT provides services including account management, task management, participant control, and participation payment transaction. However, since Amazon Mechanical Turk does not reveal information about the participants and all tasks were completed over the Internet, we had no knowledge of the background of the workers nor the quality of their work. To control the quality of the collected labels and to prevent the use of scripts or robots that can automatically fill out web forms, we embedded mechanisms and checkpoints in the interface as well as in the submission approval process:

(1) On the front of the experiment page, we specified the hardware and software requirements, and provided instructions and links to help people set up for the study (Figure 6-9(a) and (b)).

(2) Before subjects could proceed to the actual study, there was a login page with auditory captcha of a person reading a sequence of letters and numbers. Subjects were required to enter what had been said correctly in order to access the experiment page (Figure 6-9(c)). This step ensures that people can hear the sound properly and listen carefully, avoiding a situation where “robots” hack into the system.

(3) At the beginning of each HIT, an instruction clip was played, demonstrating what kind of sound would be played, and how to answer the three questions. Participants were asked to put down mandatory words at specified places as a practice. This step serves to further check the sound system and to avoid automatically generated and thus invalid responses; it also helps participants to familiarize themselves with the interface and gives an idea of the desired level of description detail.

(4) Once labels were submitted, our system compared the new results with the ground truth data from the pilot study to ensure that people were actually paying attention to the study and that meaningful labels were assigned. Finally, a manual review determined whether to accept or reject the work.



HearMe: tag special effect sound

Welcome to HearMe!

Thank you for participating in our study of sounds.

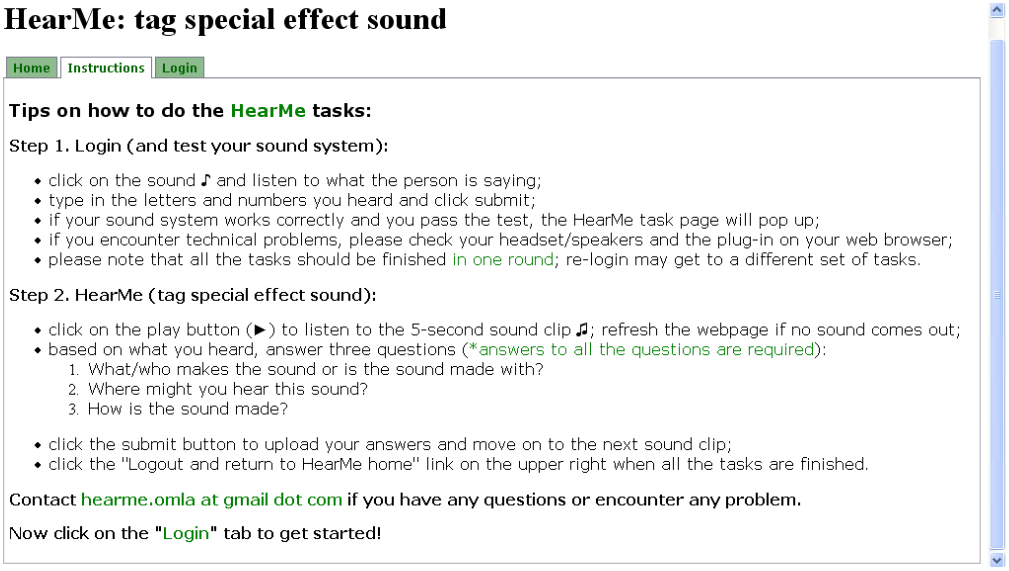
Do you have:

- a headset or speakers?
- proper plug-in on your web browser to play music? (i.e. QuickTime)

If you are ready, let's get started:

- first time users: click on the "Instructions" tab for tips on how to do the tasks;
- return users: welcome back, and click on the "Login" tab.

(a) Hardware and software examination



HearMe: tag special effect sound

Tips on how to do the HearMe tasks:

Step 1. Login (and test your sound system):

- click on the sound ♪ and listen to what the person is saying;
- type in the letters and numbers you heard and click submit;
- if your sound system works correctly and you pass the test, the HearMe task page will pop up;
- if you encounter technical problems, please check your headset/speakers and the plug-in on your web browser;
- please note that all the tasks should be finished in one round; re-login may get to a different set of tasks.

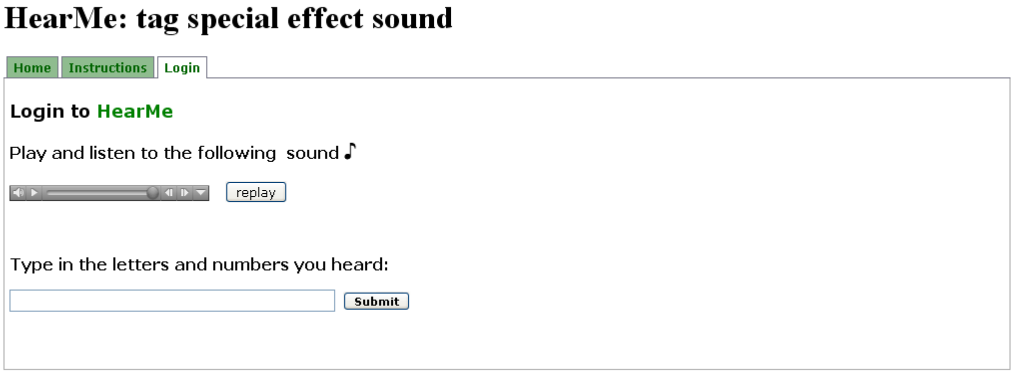
Step 2. HearMe (tag special effect sound):

- click on the play button (►) to listen to the 5-second sound clip ♬; refresh the webpage if no sound comes out;
- based on what you heard, answer three questions (*answers to all the questions are required):
  1. What/who makes the sound or is the sound made with?
  2. Where might you hear this sound?
  3. How is the sound made?
- click the submit button to upload your answers and move on to the next sound clip;
- click the "Logout and return to HearMe home" link on the upper right when all the tasks are finished.

Contact hearme.omla at gmail dot com if you have any questions or encounter any problem.

Now click on the "Login" tab to get started!

(b) Instructions



HearMe: tag special effect sound

Login to HearMe

Play and listen to the following sound ♪

Type in the letters and numbers you heard:

(c) Login and capcha

Figure 6-9. Quality controls.

In our sound labelling study, the 327 soundnails were shuffled and randomly grouped into 32 assignments of 10 to 11 sounds each, noted as Human Intelligence Tasks (HITs) by AMT. The size of the HIT was based on the response time logged in the pilot study, which avoids an overly long or tiring task. We requested at least 100 people to label each HIT, and no one person could label the same HIT twice. On average, the completion time per HIT was 14.64 minutes. Data collection took a total of 97 days. Individual completion time per sound was logged.

Although AMT does not provide any identity and demographic information of the workers, we collected the geographic location for a sample of the participants (Table 6-3). People from 46 countries took part in the experiment, which suggested that our results had universal and culture-independent validity.

country	participants	country	participants
United States (49 states)	1344	Macedonia	15
India	465	Bahamas	12
United Kingdom	49	Philippines	12
Canada	48	Germany	11
Egypt	24	Others	55

Table 6-3. Examples of country and participant counts for AMT study.

6.3.4 Data Processing

After the AMT online sound labelling study was completed, each soundnail had been labelled by at least 100 (up to 174) participants. All labels were in short phrases or sentence format. To facilitate analysis and evaluation, the human label data were processed as follows.

Each sentence was broken down into bags of words. Functional words that do not contain much information, such as “the,” “and,” etc. were filtered out. The raw data contained inflected words that we stemmed (reduced to their base forms) with the help of WordNet and the Natural

Language Toolkit (NLT). Each unstemmed word was first looked up in WordNet to see if it has a meaning independent of the base form; if this was not the case, it was stemmed. For example, “woods” meaning “forest” was not reduced to “wood,” since it has its own meaning, while “pens” was transformed back to “pen.” Following these steps, each sound was associated with a set of validated words.

For each sound, the total number of times each valid word appears was counted across all labellers. This number is referred as "**word count**" in the following sections. Table 6-4 shows the top five word counts for the three soundnails associated with target word “train.” This suggests that the concepts that people associated varied according to the particular sound scene (three in this example) even though they have the same source.

sound	**train, arrive at station**		**train, choochoo**		**train, pass on bridge**	
	word	**count**	**word**	**count**	**word**	**count**
1	train	137	train	255	train	199
2	station	31	whistle	72	track	94
3	track	30	station	69	moving	31
4	horn	22	track	42	railroad	26
5	movie	16	blow	38	helicopter	20

Table 6-4. Word count examples for the three sounds for train.

sound (top 10)	**tag length (avg.)**	**sound (bottom 10)**	**tag length (avg.)**
Beer, Plopp	9.66	Bird, Chaffinch	5.47
Ball, Table Tennis Ball	9.32	Doorbell, Single	5.45
Beer, Pour	9.24	Turkey, Hen	5.42
Reverse, Truck Backup	9.22	Wind, Howling	5.39
Zoo, Bird Dog and People	9.21	Bird, Nightingale	5.34
School, Classroom Bell	9.18	Cat, Persian Meowing	5.31
Weight, Off the Scale	8.86	Cold, Coughing	5.28
Beer, Bottle Open	8.74	Farm, Hen House	5.18
Roll, Ball	8.71	Farm, Cattle in Shed	5.12
Move, Concrete Block	8.67	Bird, Linnet	4.85

Table 6-5. Top 10 and bottom 10 sounds in their average word count.

The average number of valid words generated per person per sound was calculated. This is indicative of how much information a sound can evoke, i.e., how descriptive a sound is. Table 6-5 shows the ten most descriptive and the ten least descriptive sounds by the average number of words contained in the tags given by individual labellers. The sounds are listed as “pre-assigned word (in SoundNet), description”. For example, “Cat, Persian Meowing” is a soundnail for the word “cat,” and it renders a Persian cat meowing. It is interesting to see that nine out of the ten sounds that have the lowest word count are generated by living beings (human or animal) or natural phenomena (such as wind), suggesting that natural sounds tend to be more distinctive.

Also, more descriptive sounds (e.g., a doorbell ringing) only involve one source with no interaction among objects. In these cases, fewer words are needed to explain how the sounds were generated. By contrast, sounds with high average word counts could be associated with a more complex scene (such as zoo and school), or with several possible sources (e.g. heaviness).

6.3.5 Sound Label Sense Disambiguation

After the data processing, between four to ten words from the human labels were associated to each soundnail. However, given the polysemy of words, which concepts are considered relevant to the sounds remains unknown. It requires further work, i.e. disambiguating the senses of the label words given the sound scenes. Since the soundnail precision study generated over three thousand words, assigning meanings to each word by hand was time consuming and labour intensive. From a broader point of view, any manual approach is of poor scalability – hard for the designers and users to add new multimedia representations for concepts that are missing in the vocabulary.

Therefore, a new approach was explored for generating a vocabulary with concept to sound associations, i.e., conducting word sense disambiguation (WSD) techniques used in Natural Language Processing on the sound label datasets. For example, the labels “car, drive, fast” for the sound “car – passing.wav” are assigned to synsets “car: a motor vehicle,” “drive: operate or control a vehicle,” and “fast: quickly or rapidly” via WSD. This means that the sound “car – passing.wav” can be used to depict those concepts. This approach is viable because the words in the sound/image labels were shown to evoke one another based on the auditory/visual content, and their meanings can be identified by considering all the tags generated for a given sound or image as a context. With the availability of large sound/image label datasets, the vocabulary created from WSD can be easily expanded. The detailed algorithm description is given in *Section 4.2 Automatic AAC Vocabulary Construction from Image Label Dataset*.

Given that the accuracy of WSD cannot reach 100%, our group manually examined the concepts automatically assigned to the soundnails by the voting algorithm. Evaluation results showed that the voting algorithm achieved over 80% accuracy rate, which was significantly higher than any of the existing single-measure WSD algorithm.

6.3.6 Evaluation Metrics

Within the “bag of words” for a given sound, different words were often used to denote the same or very similar concept. In this sense, it seemed meaningful to group those words together as a **“sense set”** (or **concept group**) when considering what concepts the sound evoke. To be convenient, in the following sections, a sense set will be referred as a **“label”** to be distinguished from “word.” If not specified, all the calculations and evaluations described below are based on labels instead of words.

There are three cases how words can be grouped into labels:

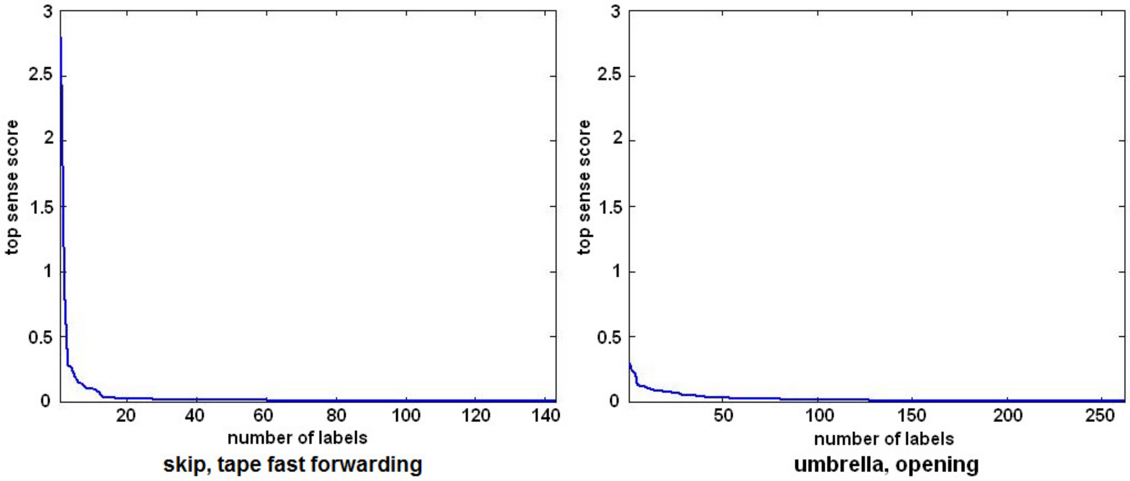
1) Synonym sets, in which words have the same meaning. For example, “baby,” “infant,” and “newborn” are grouped into the “baby” sense set, labelled with the most frequently used word “baby.”
2) Similar senses expressed by words from different parts of speech. For example, “rain (noun),” “raining (verb),” and “rainy (adj.)” are grouped into the “rain” sense set.
3) Hyponym and hypernym (super- and subordinates). This varied case by case. For example, for the sound “ball”, “basketball,” “tennis ball”, “ping pong ball” will all be put in the “ball” sense set, while for the sound “basketball”, the word “basketball” had its own sense set.

A weight is calculated for each member word in the sense set based on their actual word count. In this process, misspelled words were corrected and taken into account.

Since a word count depends on the number of participants who labelled the sound and thus varied across sounds, a relative score, referred as “**Sense Score**” was calculated for each sense set per sound. It is the average number of times across all labellers with which a sense set (label) is generated for a sound. Thus, the sense score shows how much participants agree on a label.

sense score = word count / number of labelers

For example, a score of 0.5 means 50% of the participants generate the label (sense set) once, and a score of 2 means on average each person used the label twice. The estimate of the highest sense score is 3, meaning that each person used the label once for answering each of the three questions. Figure 6-10 is an example of sorted sense score distribution for a sound.



skip, tape fast forwarding **umbrella, opening**

Figure 6-10. Examples of a sorted sense score distribution. Each point on the x- axis is a label (sense set); the corresponding y-axis value is the sense score for the label.

Based on the sense score distribution, statistical descriptors can be calculated. The evaluation metrics (non-statistical and statistical) we looked at in our primary analysis include:

(1) **Top score**: the highest sense score for each sound. It shows the extent to which participants agreed on the most agreed-upon word. It is the highest y-axis value in a sense score distribution plot.

(2) **Label number**: the total number of labels (sense sets) generated for each sound. It is the highest x-axis value in a sense score distribution plot.

(3) **Mean score and standard diviation**: mean and stdev of the sense scores.

(4) **Steepness**: this measure shows quickly the sense scores drop across labels. Usually, the flatter the sense score distribution is, the less clearly the sound is associated with a single concept.

(5) **Entropy**: this is another measure of sense score distribution, defined as the formula below (p_i is the sense score for the i^{th} label for the given soundnail).

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

(6) **Response time**: the time (logged by interface) between the sound starting to play and the participant submitting his/her answers. Although the response time could be affected by factors like how quickly a sound is perceived, how long a sentence was used to describe the sound, how fast he/she typed, etc., it still can reflect whether or not people had difficulty identifying a sound.

Figure 6-11 shows the histograms for top score (left) and label number (right) across all sounds. These suggest that most of the sounds conveyed a certain concept (over 300 sounds have a top score greater than 0.5, meaning that over half of the people use the same label).

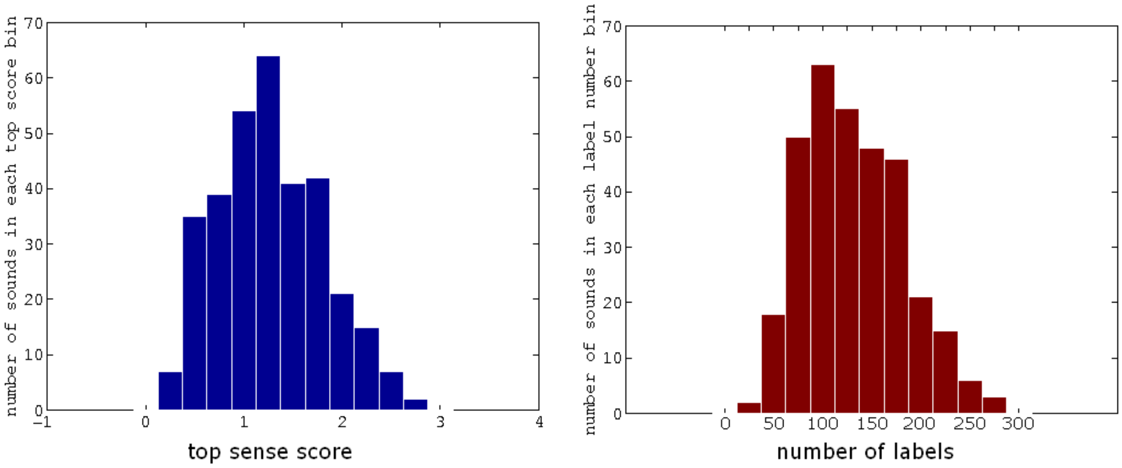


Figure 6-11. Histogram of top sense score and number of labels across all soundnails.

Correlation coefficients between the the metrics were computed. The correlation (Figure 6-12 top right) between top score and label number is about -0.7, which means in general, the more people agree on a concept, the fewer labels are generated. However, it does not neccessary mean that people use less words to describe a sound that they are sure of (Figure 6-12 top left). Figure 6-12 also shows that there is a strong correlation (> 0.9) between top score, mean score, standard deviation, and steepness. Figure 6-13 shows that the more distinctive a sound, the more people agree with one another (entropy), and the less time was needed for people to listen and respond (response time). The correlation between response time and top sense score is -0.55, while response time and tag length has a correlation of 0.40. It suggests that response time was affected by both the distinctiveness of the sound and the descriptiveness of the human labels. Based on the analysis above, we simply used the top score and response time for evaluating how well a sound can convey a concept in the rest of the discussion.

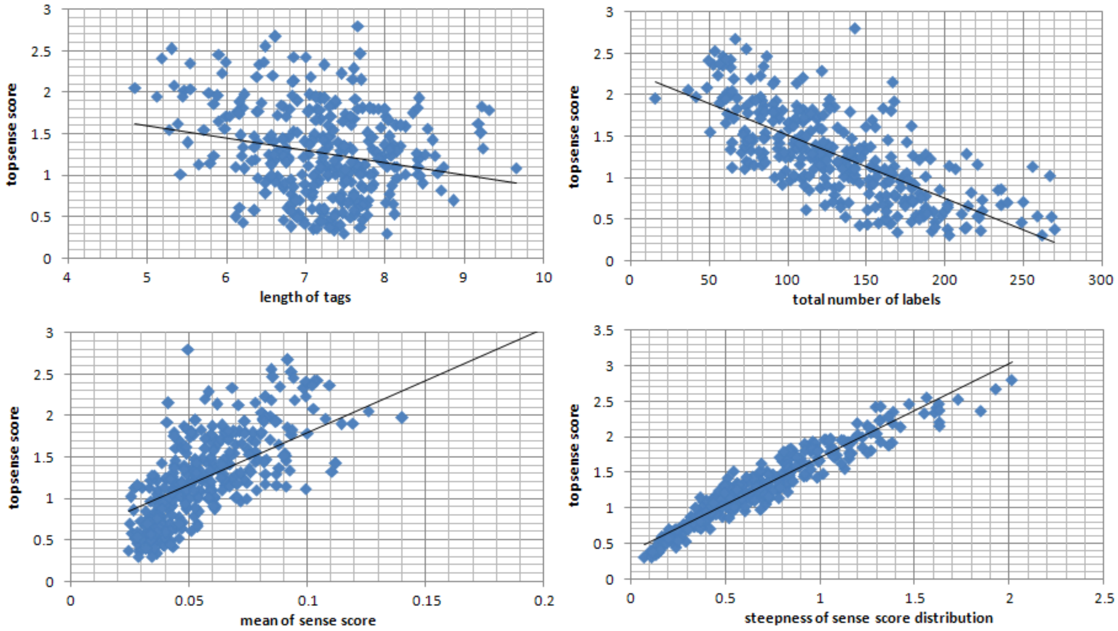


Figure 6-12. Correlation between top score and tag length (top left), label number (top right), mean sense score (bottom left), and sense score steepness (bottom right).

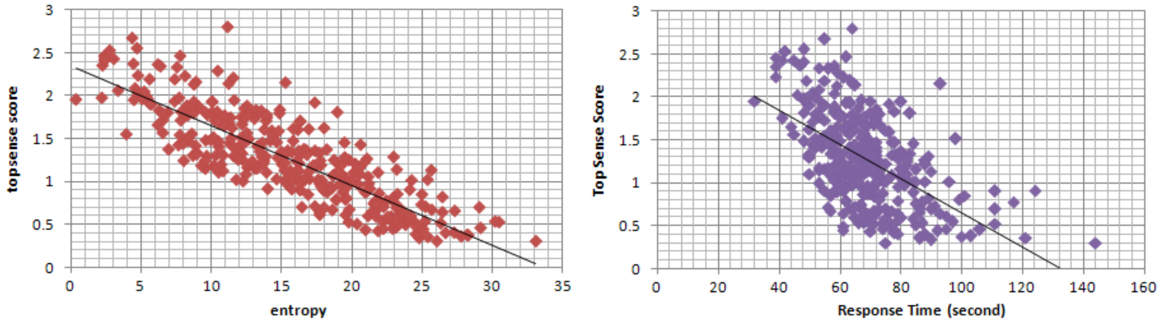


Figure 6-13. Correlation between top sense score and entropy (left) and response time (right).

The following results and analysis sections are organized around the research questions proposed in *Section 6.3.1*. First, concept level analysis on various linguistic properties generates an understanding of what kinds of concepts can be expressed through environmental sounds. Second, sound level analysis allows us to look at the evocativeness of different classes of sounds and the influence of various acoustic characteristics. Third, analysis on confusion errors provides deeper insight into the perception process of human perception of environmental sound as well as the semantic structure embedded in the auditory events. Fourth, information drawn from the limited knowledge about the anonymous AMT labellers suggests some interesting findings. At the end, a general discussion introduces other factors that may affect people's perception of information based on non-speech auditory representations.

6.3.7 Results and Analysis, Concept Level

This section inspects the soundnail labelling data from the linguistic perspective, targeting the question of what concept types have high audioability (are illustrated well by non-speech audio).

6.3.7.1 Pre-assigned Concept vs. Most Agreed-on Concept

Since a major goal of our study is to determine whether the soundnails can illustrate the concepts pre-assigned in SoundNet, the sense scores of intended sense set were extracted and compared to the sense sets showing the highest agreement among the participants. The results can be categorized into four kinds of situations, exemplified in Table 6-6:

(1) For those sounds whose pre-assigned concept shows the highest agreement, the results confirm that they successfully convey the intended concept. There are about ninety sounds in this category. These soundnails are effective and can likely be utilized to assist language comprehension and communication.

(2) For the sounds of which the label with the highest agreement (different from the pre-assigned concept) matches the sound description (given in the sound caption), it can be said that the sound (scene) is distinctive and can convey a concept, though different from what is desired. There are 150 sounds in this category. Two possible reasons can be cited for this result. (a) The desired concept requires extra linkage to the sound scene;(b) the participants focused on different objects or aspects related to the sound event.

(3) The sounds where participants provided labels different from the sound description with high agreement are suggestive of a concept, though not the a priori one. 52 sounds fall into this category.

(4) In the case of the sounds for which participants in general did not agree (low top scores), we conclude that they lack the necessary characteristics for people to identify and associate them with specific concepts. There are 35 sounds within this category.

Of course, cases (2- 4) may simply suggest problems with the scripting and sound selection. Further analysis on why people came up with different labels than what was desired can guide our future refinement of the construction of a network of concept-nonlinguistic audio connections (see *Section 6.3.9 Results and Analysis, Confusion Errors*).

case	sound	pre-assigned concept	most agreed-on label
1)	Cat, Persian Meowing	cat	cat
1)	Train, Choochoo	train	train
2a)	Farm, Cattle in Shed	farm	cow
2a)	Day, Rooster Clock Crickets	day	rooster
2b)	Floor, Walk in Classroom	floor	walk
2b)	Toilet, Flush	toilet	water
3)	Television, Change Channel	television	radio
3)	Slice, Cut Bread	slice	saw
4)	Umbrella, Open Umbrella	umbrella	match
4)	Bike, Wheel Turn	bike	motor

Table 6-6. Examples of situations of how well sounds convey target concepts.

6.3.7.2 Audioability Score vs. Actual Auditory Evocativeness

As one may recall, when building SoundNet, an audioability rating was assigned to each concept in the inventory by human raters (*Section 6.2.2 Audioability Rating*). This audioability score is a new linguistic property that measures how well a concept can be conveyed via an environmental auditory representation. For each soundnail, I compared the audioability rating and sense score of its pre-assigned concepts in SoundNet to those of the most agreed-on label obtained in the AMT study. Table 6-7 shows the top five and bottom five soundnails based on the sense scores of pre-assigned concepts. A test for homogeneity of variances showed that sense scores for the pre-assigned concepts and most agreed-on labels came from the same normal distribution. It suggests that if the intended concepts are strongly audioable (with a rating 3 as indicated in the parentheses), they are likely to be agreed-upon by labelers. On the contrary, people tend to come up with a different and more audioable concept if the pre-assigned one is less evocative.

Comparison of the audioability ratings and sense scores of the target concepts is shown in Figure 6-14. ANOVA shows that strongly audioable (rating 3) concepts received a significantly higher sense score, and scores for non-audioable concepts were significantly lower ($F(1, 206) = 19.941$, $p < 0.001$, $\eta^2 = 0.088$). Sense scores for concepts with audioability rating 1 and 2 do not quite follow the trend. It is because it is hard for people to judge whether they can identify a sound with borderline distinctiveness.

sound	pre-assigned			most agreed-on		
	concept	audioability	S.S.	concept	audioability	S.S.
cat_meowing	cat	3	2.53	cat	3	2.53
train_choochoo	train	3	2.46	train	3	2.46
telephone_ring	phone	3	2.43	phone	3	2.43
horn_carHorn	horn	3	2.42	horn	3	2.42
baby_happy	baby	3	2.36	baby	3	2.36
empty_waterOut	empty	2	0	water	3	1.68
teapot_waterFill	teapot	1	0	water	3	1.71
speed_carTurnFast	speed	2	0	car	3	1.71
skip_tapeForward	skip	1	0	projector	3	1.81
cracker_eatCrunch	cracker	2	0	eat	3	1.91

Table 6-7. The five most and least effective soundnails with audioability ratings and sense score (S.S.) for their pre-assigned concept and the most agreed-on labels.

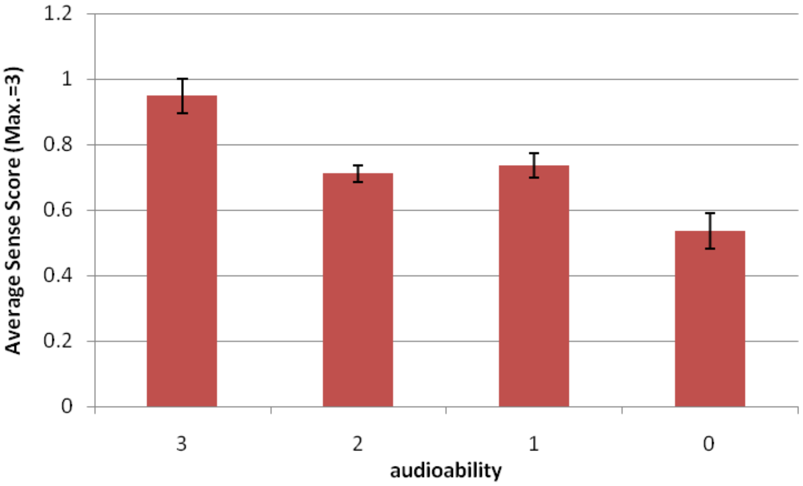


Figure 6-14. Comparison of audioability rating and sense score.

6.3.7.3 Parts of Speech

Evaluations of visual representations have shown that part-of-speech has a strong effect on the potential of a concept being effectively visualized. Same analysis was conducted on non-speech auditory representations to investigate the expressiveness differences across different parts of speech. Table 6-8 compares the numbers of words from different parts of speech for the target words and the most agreed-upon label for all sounds. The table further lists the parts of speech distribution among words (broken down the sense sets) that at least 25% of the participants agreed to assign to a sound. Table 6-9 gives further details in differences between intended parts of speech and labelled parts of speech. Note that the total number does not add up to 327 because some words can be in different parts of speech though with similar meanings. For example, people used “rain” both as a noun and a verb in their labels. Difference in labels by part of speech here does not impact their audioability rating. About 80% of soundnails intended for a noun were labeled as a noun, while only half of those for a verb and almost all for adjectives and adverbs were most agreed upon as nouns.

POS	pre-assigned words	most agreed-on labels	most agreed-on words
Noun	250	271	1328
Verb	98	65	488
Adj.	14	4	117
Adv.	6	0	31

Table 6-8. Comparison of numbers of different parts of speech in target words and most agreed labels for all sounds.

POS, intended	POS, agreed	count	POS, intended	POS, agreed	count
Noun	Noun	231	Adj.	Noun	14
	Verb	56		Verb	2
	Adj.	4		Adj.	2
	Adv.	0		Adv.	0
Verb	Noun	38	Adv.	Noun	6
	Verb	39		Verb	1
	Adj.	0		Adj.	0
	Adv.	0		Adv.	0

Table 6-9. Pairwise comparison between parts of speech of the target words and those of the most agreed-upon labels.

Figure 6-15 shows the sense score for intended (pre-assigned) concepts and the most agreed-on labels for different parts of speech. Results showed that it was significantly more likely for people to generate a noun than a verb, and even more than an adjective or adverb (for pre-assigned concepts: F(3,204) = 3.296, p = 0.022, η^2 = 0.767). Audioability ratings across parts of speech have the same trend (F(3,204) = 3.735, p = 0.012, η^2 = 0.052).

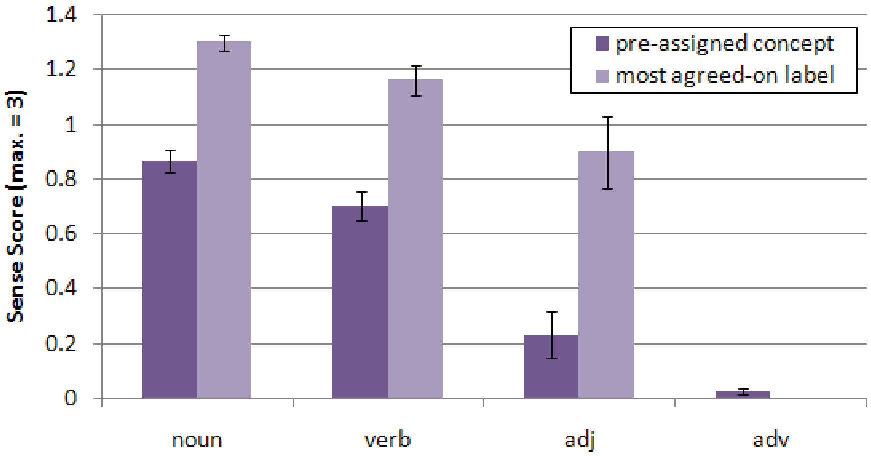


Figure 6-15. Comparison of sense score of pre-assigned concepts and most agreed-on labels in different parts of speech.

The response time and length of tags, given the different parts of speech that the most agreed-on labels were in, were compared (Figure 6-16). ANOVA results showed that it took significantly longer time for people to agree on an adjective based on what they perceived from the environmental sound than on a noun or a verb ($F(2, 363) = 3.390$, $p = 0.035$, $\eta^2 = 0.018$). More (but not significantly) words were used to describe sounds that the listeners considered depict adjectives than those for nouns and verbs ($F(2, 363) = 0.563$, $p = 0.570$, $\eta^2 = 0.003$).

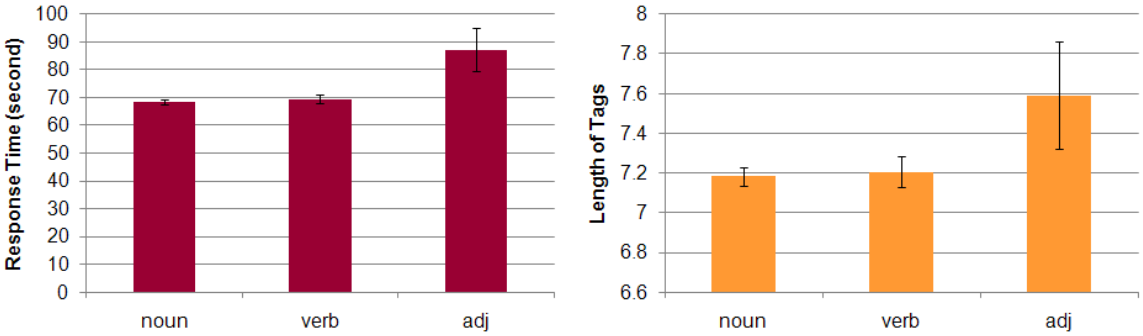


Figure 6-16. Comparison of response time (left) and length of tag (right) given the most-agreed label in different parts of speech.

To better understand what kinds of words (objects, places, or actions) and parts of speech (nouns, verbs, adjectives, and adverbs) can be easier and better conveyed by a sound, we separated the labels by which question they were answering (what, where, and how), and ran analyses individually.

Figure 6-17 shows that more words and sense sets were generated in answering how a sound could be made. This is because longer sentences are needed to describe an interaction (identifying the sources and depicting the actions) while to answer “what” and “where”, people basically just need to name the object(s) and place(s). The differences on the types of question were significant (Table 6-10).

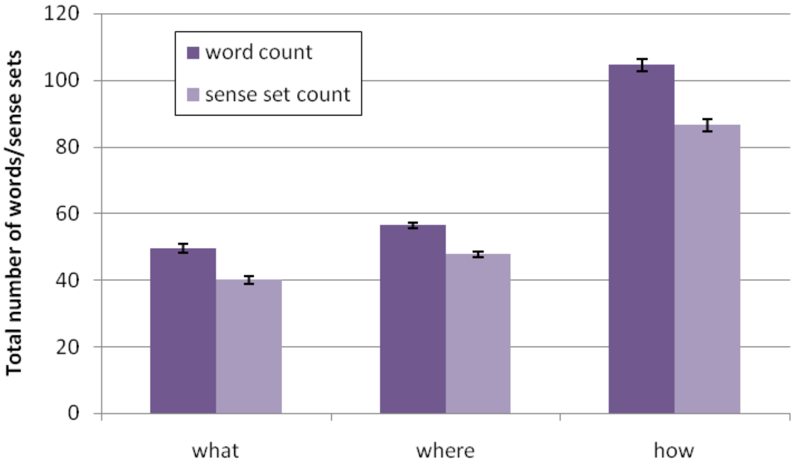


Figure 6-17. Comparison of word count and sense set count among answers to the three questions.

	word count			label count			top sense score		
	F(2,978)	p-value	η^2	F(2,978)	p-value	η^2	F(2,978)	p-value	η^2
question	424.85	<0.001	0.465	331.84	<0.001	0.404	67.668	<0.001	0.122

Table 6-10. ANOVA results on word count, label count, and top sense score differences among answers to the three questions.

Although more words were used, Figure 6-18 shows that people are more likely to agree on what kind of actions were involved than what object(s) it was and where it took place. Further, it is even harder for people to recognize the location than the source, partially because some sound can be heard anywhere and it is hard for people to specify, for example, a human breathing sound. Among the 327 soundnails, for 195 of them people agreed on “how the sound is made” (with the highest sense score), for 76 of them people agreed on “what is the source of the sound,” and for 56 of them people agreed on “where are you likely to hear the sound.” The difference in level of agreement regarding the perception of sound source(s), location(s), and (inter)action(s) was significant (Table 6-9).

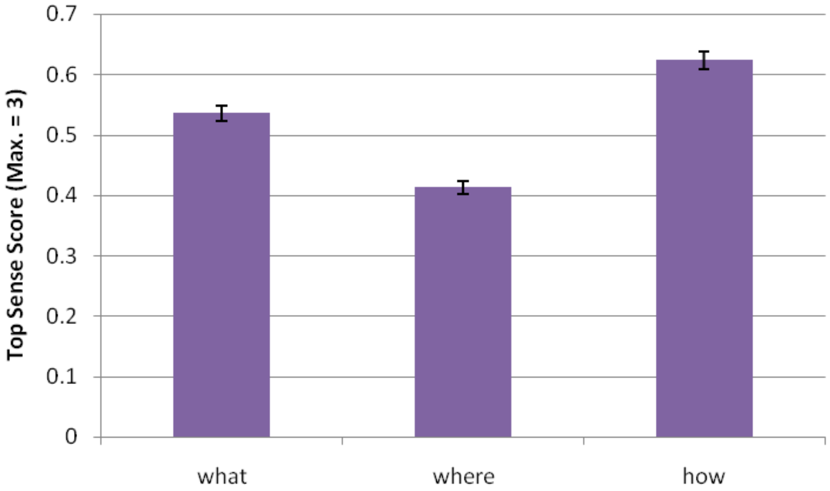


Figure 6-18. Comparison of top sense score among answers to the three questions.

Taking a closer look at the parts of speech of the most agreed sense sets in answers to the three questions (Table 6-11), it seems people came up with a lot more verbs when describing how a sound can be made, because verbs are commonly associated with actions and motions. Also, some adverbs showed up in the most agreed words for locations, because they are related to positions and directions. Table 6-12 shows the top 10 most agreed sense sets for what, where and how. Phone rings and babies are the most identifiable sound sources; farms and barns are locations that people can recognize relatively easily; and actions like zipping, printing/typying are familiar and more distinctive for people.

POS	What (source)	Where (location)	How (interaction)
Noun	313	323	256
Verb	56	15	134
Adj.	3	2	2
Adv.	0	8	0

Table 6-11. Comparison of top sense score among answers to the three questions.

	What	What POS	Where	Where POS	How	How POS
1	phone	noun	sport	noun	book	noun
2	phone	noun	road	noun	start	verb
3	baby	noun	pool	noun	typewriter,type	noun,verb
4	chicken	noun	farm	noun	zipper,zip	noun,verb
5	horn,honk	noun,verb	house	noun	paper	noun
6	doorbell	noun	road	noun	zipper,zip	noun,verb
7	baby	noun	airport	noun	horn,honk	noun,verb
8	cat	noun	school	noun	ring	noun,verb
9	bird	noun	farm	noun	printer,print	noun,verb
10	baby	noun	farm	noun	eat	verb

Table 6-12. Top ten most agreed sense sets and their parts of speech for the three questions.

Generally speaking, noun is the dominate part of speech people use to describe a sound. People have high agreement on certain verbs as well. Comparablely, adjectives and adverbs are hard to illstrate by a sound. However, if the response is directed to specific aspect of the sound events, the distribution parts of speech generation from environmental sounds can be changed.

6.3.7.4 Linguistic Properties: Concreteness, Imageability, Familiarity, and Meaningfulness

Research (Van Hell and De Groot1998) and the studies described in *Chapter 4* and *Chapter 5* suggested that concrete words and words that are highly imageable are easier to name and categorize based on pictorial representations than abstract words. Figure 6-19 (a)(b) show that concept recall via auditory representations followed the same rule. Sense score dropped significantly as concreteness and imageability (based on the MRC Psycholinguistic Database, UWA2009) went down for both target words and most agreed labels (Table 6-13). This indicates that, in general, concrete concepts and concepts that can be easily illustrated by a picture are more likely to be conveyable by an environmental sound. No significant effect was shown in the interaction between concreteness/imageability and parts of speech, suggesting that both concrete/highly imageable nouns and verbs are easier to represent via sound. Other linguistic properties like familiarity and meaningfulness had no significant impact (Table 6-13).

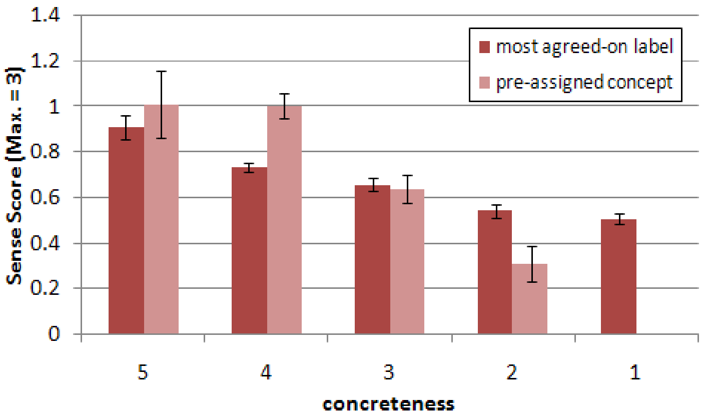


Figure 6-19 (a). Comparison of sense score of pre-assigned concepts and agreed-on labels at different concreteness level.

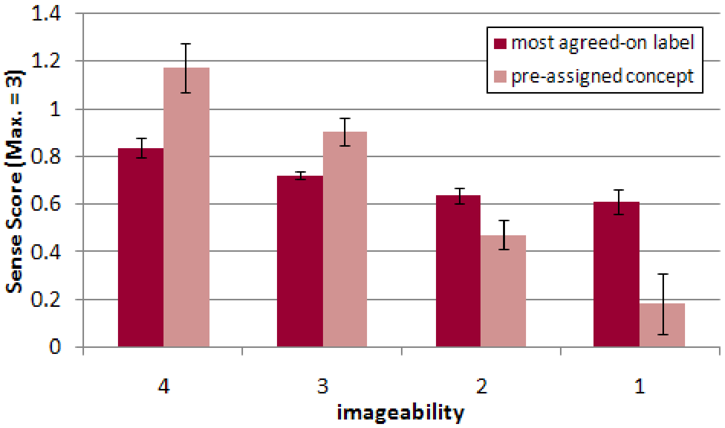


Figure 6-19 (b). Comparison of sense score of pre-assigned concepts and agreed-on labels at different imageability level.

	concepts	Df	F	p-value	η^2
concreteness	pre-assigned	1, 221	25.598	< 0.001	0.962
	most agreed-on	1, 702	33.596	< 0.001	0.971
imageability	pre-assigned	1, 221	36.062	< 0.001	0.973
	most agreed-on	1, 731	21.184	< 0.001	0.955
familiarity	all	1, 738	1.163	0.281	0.002
meaningfulness	all	1,513	2.167	0.142	0.004

Table 6-13. ANOVA results on concreteness, imageability, familiarity, and word frequency for pre-assigned concepts and most agreed-on labels.

6.3.7.5 Semantic Relations among Concepts Perceived from Everyday Sounds

I further looked into the semantic relations (based on WordNet) between sense set members for each question. This gives an insight on the possible causes for confusion, discussed later in this chapter. The semantic relationships discussed here includes synonyms (e.g. car-auto mobile), hypernyms (e.g. vehicle-car), hyponyms (e.g. sports car-car), meronyms (e.g. car window-car), holonyms (e.g. window-windowpane), sisters (e.g. truck-car), nephews (e.g. fire truck-car), and instances (e.g. Ford car-car). Table 6- shows that over 1/3 of the words in the responses to each question are synonyms to the representative word for the sense set they belong to, around 10% are hyponyms. However, hypernyms and meronyms receive relatively higher scores (bold in Table 6-14). This suggests that people are more likely to recognize a more generic scope of the actual source, location, and action in the sound, or detect part of them. People usually got confused with objects or interactions in the sister or nephew categories, and even with completely unrelated events that cause similar effects or generate similar sounds.

question	response semantic relation	percentage	sense score
what	**synonym**	**39.95%**	**0.2085**
	hyponym	10.26%	0.0354
	sister term	6.91%	0.0411
	hypernym	**6.76%**	**0.0532**
	similar sound	3.86%	0.0358
	nephew term	3.38%	0.0270
	meronym	**3.31%**	**0.0829**
	instance	2.86%	0.0424
	holonym	1.93%	0.0324
where	**synonym**	**33.86%**	**0.1817**
	hyponym	9.74%	0.0506
	hypernym	**8.76%**	**0.0933**
	meronym	**7.88%**	**0.0448**
	nephew term	4.79%	0.0429
	sister term	4.35%	0.0487
	similar place	4.11%	0.0392
	instance	2.40%	0.0307
	holonym	2.10%	0.0588
how	**synonym**	**34.90%**	**0.2412**
	hyponym	9.00%	0.0430
	sister term	6.38%	0.0478
	hypernym	**5.17%**	**0.0637**
	similar sound	4.59%	0.0521
	nephew term	3.98%	0.0396
	meronym	**3.91%**	**0.1053**
	instance	3.15%	0.0378
	similar effect	2.37%	0.0444
	holonym	2.04%	0.0331

Table 6-14. Semantic relations between sense set members.

In summary, concepts that are considered highly audioable, concrete, and imageable are more likely to be effectively evoked by an environmental auditory representation. Nouns and verbs are easier for people to generate from perceptive auditory events. The hypernyns (superordinates) and meronyms (parts) of an audioable concept are likely to be audioable.

6.3.8 Results and Analysis, Sound Level

After examing what kinds of concepts are audioable, I explored the data from a different angle, i.e. the evocativeness of environmental sounds.

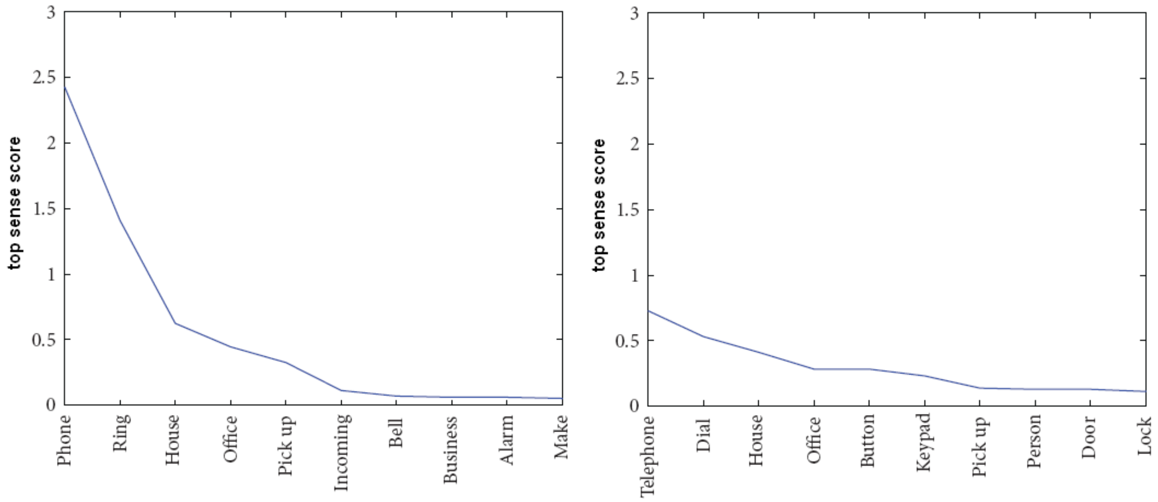
6.3.8.1 Audio Expressiveness

Table 6-15 lists the ten sounds with the highest top score and the ten sounds with the lowest top score. A high top score means people have a very similar idea on what the sound is about, while a low top score means people have very different opinions. The former indicates that sounds are distinctive and can easily be associated with a concept. The latter indicates that sounds are ambiguous and can easily be associated with more than one sound.

Sound (top ten)	Top score (label number)	Sound (bottom ten)	Top score (label number)
Skip, Tape	2.8000 (143)	Spring, Door Spring Vibrate	0.4000 (194)
Ring, Telephone	2.6716 (67)	Stop, Hose Pipe	0.3966 (214)
Scream, Baby	2.5524 (74)	Cold, Teeth Chatter	0.3814 (213)
Cat, Persian Meowing	2.5254 (54)	Bucket, Throw Can into Bucket	0.3727 (202)
Train, Choochoo	2.4622 (87)	Window, Window Slide Open	0.3712 (270)
Cry, Baby Girl Cry	2.4608 (59)	Gym, Intensive Workout Breathing	0.3551 (223)
Telephone, Ring Pick Up	2.4274 (61)	Bike, Wheel Turning	0.3500 (192)
Horn, Car Horn	2.4224 (64)	Ball, Croquet Hit	0.3363 (170)
Farm, Hen House	2.4123 (50)	Dryer, Hairdryer Stop	0.3008 (203)
Young, Baby Talk	2.3697 (57)	Umbrella, Opening Umbrella	0.3000 (262)

Table 6-15. Top ten and bottom ten sounds in top sense score.

One of our hypotheses, that different sounds that can be associated with the same concept may differ with respect to the effectiveness with which they evoke the concept, was verified. For a first test of this hypothesis, we assigned different sounds to most of the target words. For example, for “microwave,” both the sounds that were pre-labelled “loading dishes and close the door” and “spinning and ending” were given. Figure 6-20 shows the sense score distribution of the top ten sense sets for the two sounds assigned to “telephone.” One can see that the “phone ringing” sound is more distinctive than the “lifting receiver and dialling” sound. The differences for the agreed-upon sense sets for the two sounds also suggested that different scenes, events, and actions could be evoked even with the same sound source.



(a) Telephone, Ring and Pick up **(b) Telephone, Lift Handle and Dial**

Figure 6-20. Top 10 most agreed-on labels for two soundnails for the concept “telephone”.

6.3.8.2 Environmental Sound Classification

To better assess what kind of sounds are more recognizable, the 327 soundnails were classified according to three criteria reflecting the three questions our study focused on. The classification categories were suggested by a linguist and a sound specialist.

(1) Source: the source of the sounds (Table 6-16). We made a broad initial cut, distinguishing sounds involving people, sounds involving animals, sounds made by artifacts and sounds occurring in natural settings. More-fine grained distinctions involved separating the sounds generated by mechanically operated artifacts from those of electronic devices.

(2) Scene: location where the sounds are likely to take place (Table 6-17). We distinguished sounds occurring in indoor and outdoor settings. More specifically, indoor sounds are characteristic of homes, schools, shops, and offices. Outdoor sounds can pertain to activities like walking or sports, or sounds associated with natural phenomena like wind and rain.

(3) Event: complexity of interaction participants involved in the sound (Table 6-18). Finally, we classified the sounds according to their complexity as measured in terms of number of participants. Sounds can be generated by single source, such as wind, coughing, or a bell ringing (note that these cut across the categories in the first classification). Other sounds are produced by two object interaction or a human manipulating an object. Finally, very complex sounds involve multiple participants.

Source	Description
HumanVocal	Vocal sounds made by humans, e.g. coughing and laughing.
HumanContact	Actions initiated by human e.g. walking on the snow and knocking on the door.
HumanScene	Complex scenes that involve humans, e.g. a football game.
NaturalAnimal	Sounds made by animals, e.g. birds and crickets.
NaturePhenomenon	Natural sounds not made by animals, e.g. wind and waves.
NatureSemi	Artificial sounds similar to natural ones e.g. human blowing air/splashing water.
ObjectContact	Sounds made by contact between two objects, e.g. a bat hitting a ball.
ObjectRustle	Sounds made by object rustling, e.g. rustling a plastic bag.
ToolVehicles	Sounds related to cars, boats, and planes as well as their parts.
ToolMechanical	Sounds made by mechanical tools, e.g. scissors and handsaw.
ToolMachine	Sounds made by machine or electric device, e.g. a washer or drill.
ToolElectronic	Electronic devices such as television and radio.
BellAlarm	All kinds or alarms and sirens.
BellOrdinary	Ordinary (non-electronic) bells, e.g. doorbells and church bells.
BellSynthetic	Synthesized/synthetic sounds.

Table 6-16. Descriptions of sounds divided by category.

Scene	Description
Outdoors	Sounds taking place outdoors but cannot be specified, e.g. wind
Indoors	Sounds taking place indoors but cannot be specified, e.g. step on the floor.
Bathroom	Sounds taking place in the bathroom, e.g. flushing.
Kitchen	Sounds taking place in the kitchen, e.g. washing dishes.
School	Sounds taking place in the school, e.g. classroom.
Office	Sounds taking place in an office, e.g. printing.
Workshop	Sounds taking place in a workshop/factory, e.g. hammering.
Transportation	Transportation-related sounds, e.g. car sounds and stations.
Sport	Sports-related sounds, e.g. basketball games and jogging.
Transaction	Transaction-related sounds, e.g. a cash register.
Nature	Sounds happening in nature scenes, e.g. birds singing.
Uncertain	Sounds that can take place anywhere, e.g. coughing.

Table 6-17. Descriptions of sounds divided by Scene.

Event	Description
SingleSource	Sounds initiated and completed by a single source, can be divided into finer groups: SingleNature and SingleArtificial.
SingleNature	Single source sounds made by living beings or nature phenomenon.
SingleArtificial	Single source sounds made by bells, machines, and artificial objects.
HumanObject	Sounds of human manipulating one object, such as rustling a bag.
ObjectObject	Sounds of two objects interacting, such as pen scratching paper.
MultipleObjects	Complex sound scenes or sounds with multiple objects involved.

Table 6-18. Descriptions of sounds divided by event.

6.3.8.3 Evocativeness of Sounds from Various Sources

The difference in entropy is significant for sounds in different categories (F(14,312) = 9.125, p < 0.001, η^2 = 0.291). To be more specific (Figure 6-21 top), human vocal sounds (e.g. cough, laugh) and animal sounds (e.g. dog barking and bird chirping) are the most distinctive. People can easily describe the sound with specific words. On the contrary, synthetic or electronic machinery sounds are vaguer. They can appear in different settings and maybe associated to different things, such as alarms and sound effects in video games or movies. Examples of sense score distribution of different sound source categories are shown in Figure 6-22. The similar pattern for sounds from natural phenomena (Figure 6-22 (c)) and the sounds of gases and liquid produced artificially (Figure 6-22 (d)) suggests that in some cases the acoustic characteristics of the sound play an more important role than the actual source.

The response time for people to label sounds in different categories varied significantly as well (F(14,312) = 4.537, p < 0.001, η^2 = 0.169). Non-synthetic bell sounds and human/animal vocal sounds were recognized and labelled quickly, while it took longer for people to identify objects and machines. (see Figure 6-21 bottom).

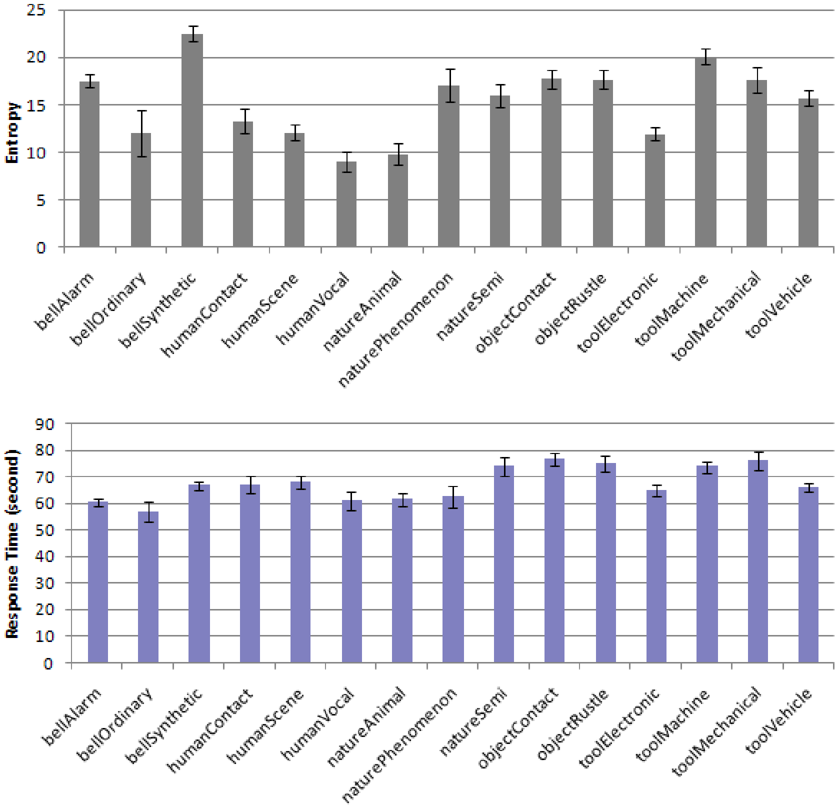
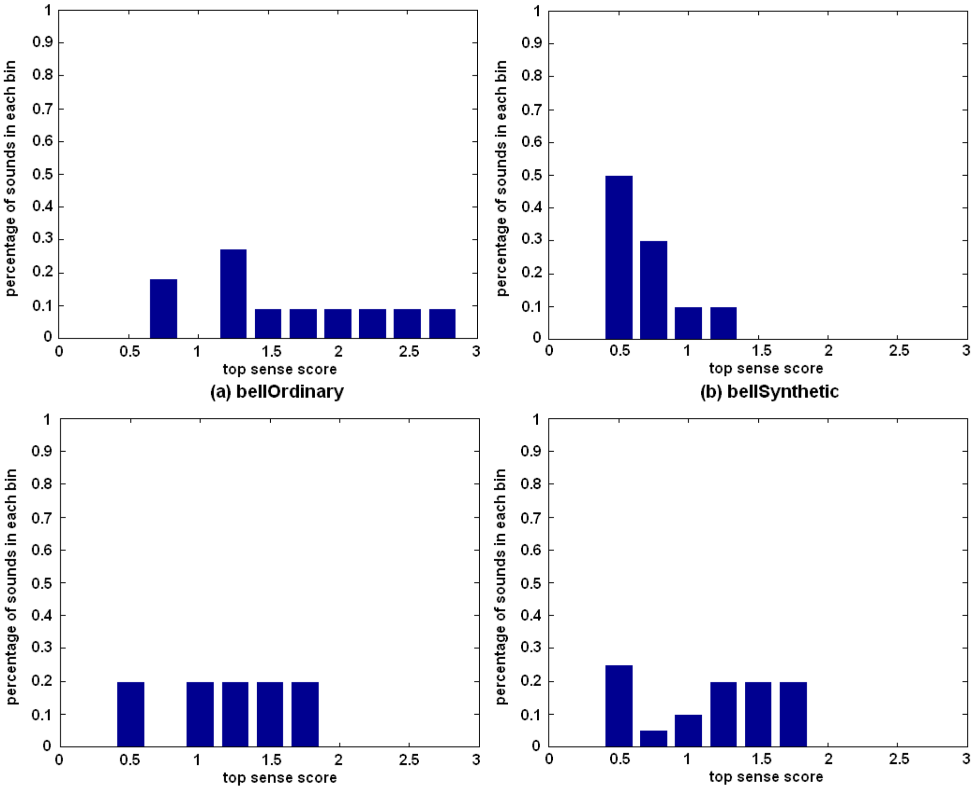


Figure 6-21. Comparison of entropy and response time across different classes of sound source.



(a) bellOrdinary

(b) bellSynthetic

(c) naturePhenomenon

(d) natureSemi

Figure 6-22. Examples of sense score distribution in different sound source categories.

6.3.8.4 Evocativeness of Sounds from Various Scenes (Locations)

When simply classifying sounds by their locations as "indoors (e.g. flushing the toilet)," "outdoors (e.g. walking in the snow)," and "uncertain (those can take place both indoor, e.g. coughing)," the differences in entropy ($F(2,324) = 3.219$, $p = 0.041$, $\eta^2 = 0.019$) and response time ($F(2,324) = 5.424$, $p = 0.005$, $\eta^2 = 0.033$) were significant. Relatively, outdoor sounds took less time to be identified, and people's responses converged more using significantly fewer words ($F(2,324) = 13.240$, $p < 0.001$, $\eta^2 = 0.076$) (Figure 6-23).

Comparisons were done on a finer classification as described in Table 6-16, and the differences were significant (for entropy $F(11,315) = 2.808$, $p = 0.002$, $\eta^2 = 0.089$; for response time $F(11,315) = 3.193$, $p < 0.001$, $\eta^2 = 0.100$). In general (Figure 6-24), people did better in recognizing school scenes (mainly "children") and nature scenes (including animals and natural phenomenon like rain and wind), and relatively worse in workshop scenes which involved various machines and tools.

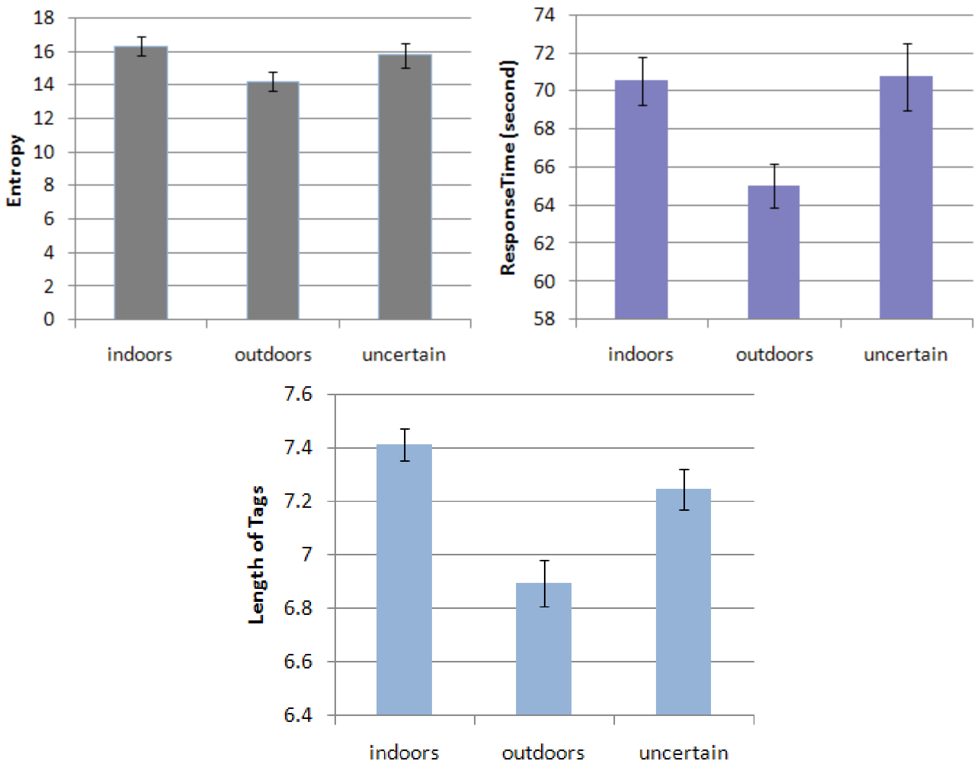


Figure 6-23. Comparison of entropy, response time and tag length across different scenes.

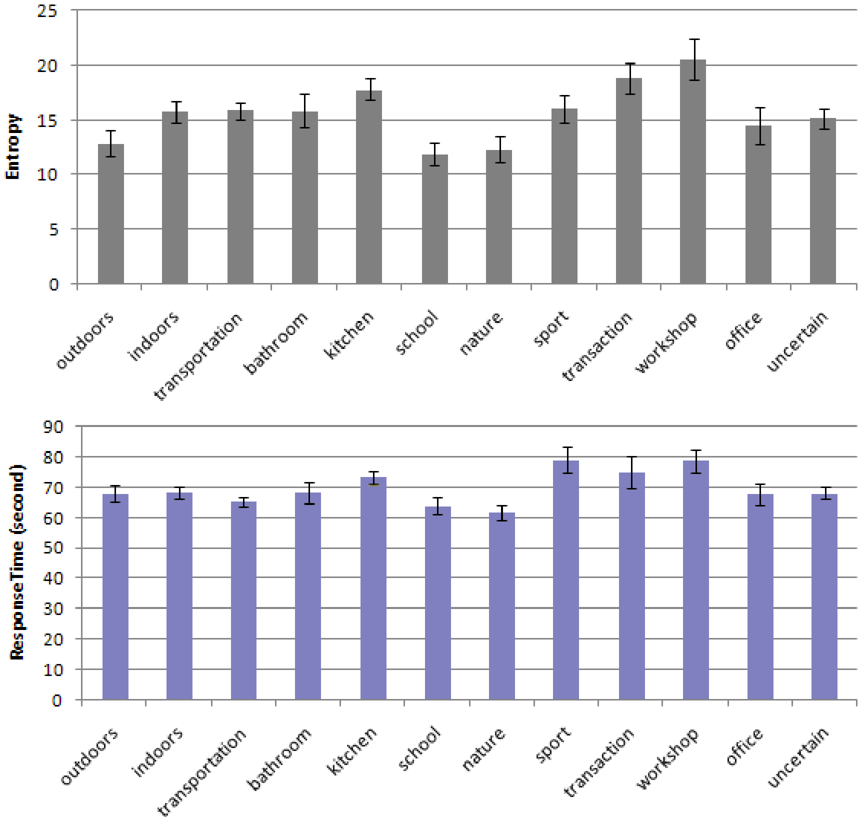
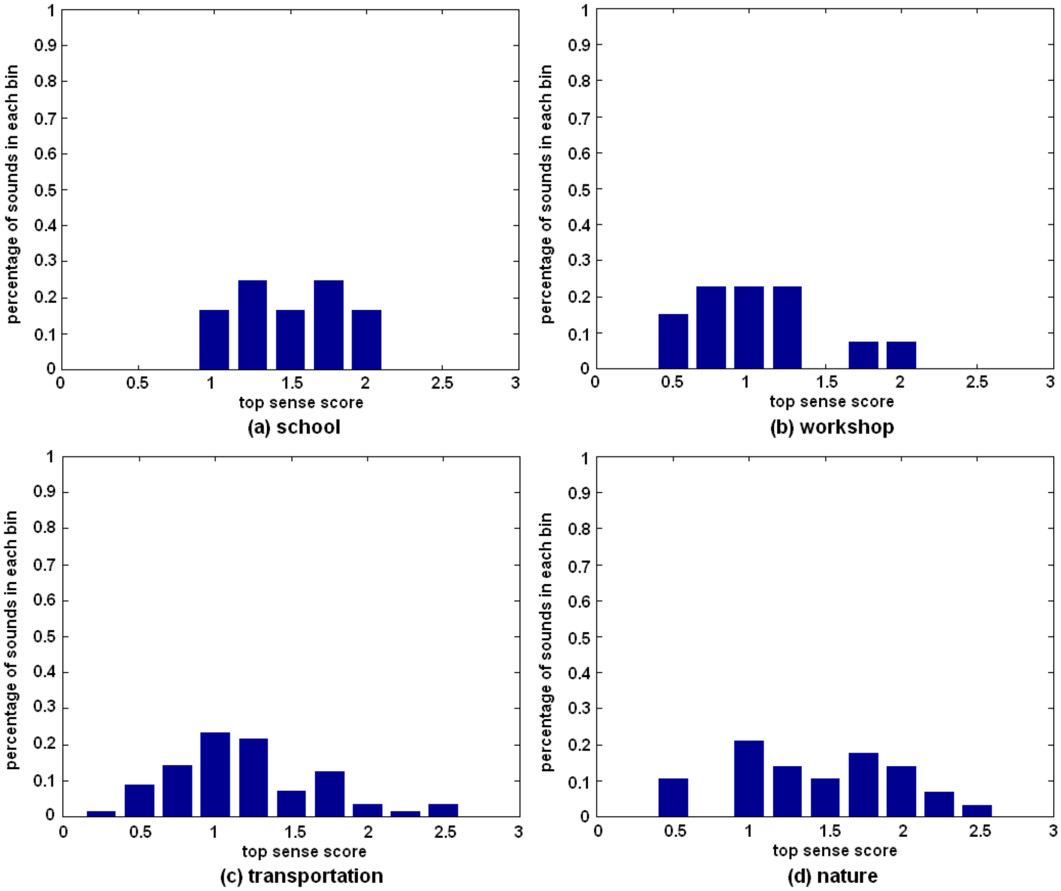


Figure 6-24. Comparison of entropy and response time across different classes of sound scenes.

Figure 6-25 gives several examples of top sense score distribution in various sound scene categories. Sounds happening in school (Figure 6-25(a)) are generally recognizable, whereas sounds related to machines used in a workshop/factory are hard to identify (Figure 6-25(b)). People are good at naming engine and horn sounds of a vehicle, but not sounds with parts of a vehicle, e.g. seatbelt (Figure 6-25(c)). A lot of the sounds occurring in natural environment are distinctive, such as wind, wave, and bird songs. However, some animal sounds are harder to discriminate, maybe due to unfamiliarity (Figure 6-25 (d)).



(a) school

(b) workshop

(c) transportation

(d) nature

Figure 6-25. Examples of sense score distribution in different sound scene categories.

6.3.8.5 Evocativeness of Sounds from Various Events (Interactions)

If looking at the sound labelling performance based on the interaction involved to produce the sound, single source natural sounds (e.g. cat, baby crying, ocean waves) were the most recognisable, while sounds coming from interaction between two objects (e.g. chopping onions, scrubbing a mattress) were more confusing (Figure 6-26 top). This difference was significant reflected in both entropy (F(4,322) = 13.060, $p < 0.001$, $\eta^2 = 0.140$) and response time (F(4,322) = 8.045, $p < 0.001$, $\eta^2 = 0.091$). Figure 26 (bottom) shows that generally the time taken to label a sound increased as the complexity of the interaction in the sound went up, except those involving multiple sources. It appears that people often focused on the most distinctive part in the complex sound scene, such as people talking and dishes clanking, instead of trying to identify everything in the scene. The top sense score distribution for sound event categories are all quite spread out, without any unique pattern.

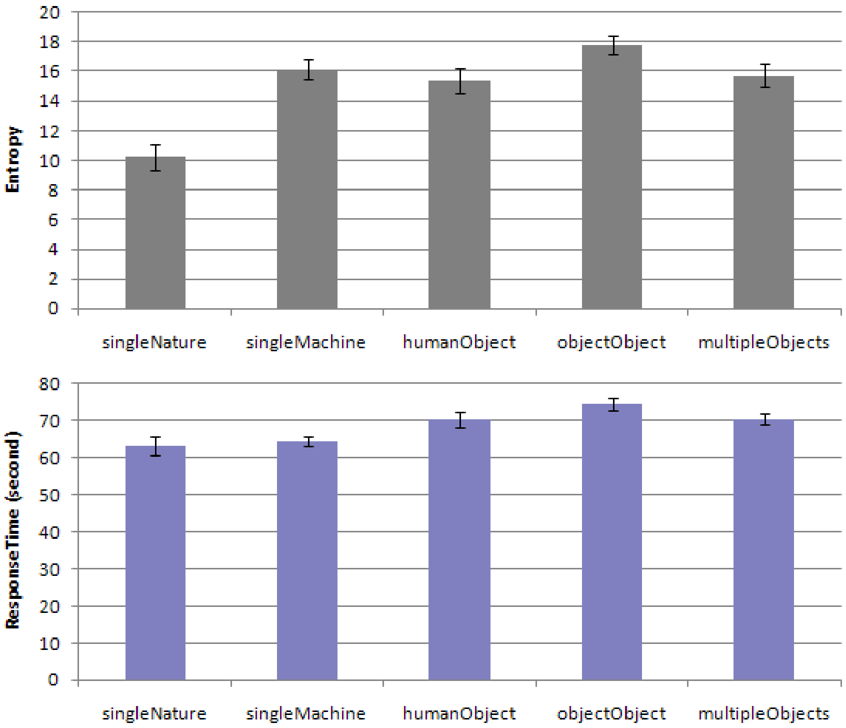


Figure 6-26. Comparison of entropy and response time across different classes of sound events.

In general, synthesized sounds are quite ambiguous. People often associate them with science fiction and movies. By contrast, single source natural sounds are very distinctive. An example of an outlier is “Cold, Teeth Chatter.” People could also identify sound scenes involving humans relatively easily. For example, they can tell a restaurant from a train station and a school from an office.

The three criteria (sources, locations, and events) for accounting for sounds’ audioability are not sufficient to explain the results. More relevant factors that impact the distinctiveness of non-speech audio should be hypothesized investigated. For example, material (glass versus metal versus stone) might be a strong indicator **Error! Reference source not found.**. A better categorization of sounds based on their expressiveness (e.g. the taxonomy introduced in *Section 6.1.5 Classification of Environmental Sounds* will provide guidance for designing improved non-speech audio representations of concepts.

6.3.9 Results and Analysis, Confusion Errors

Table 6-19 summarizes several examples of responses (some are confusion errors) people generated for the intended concepts given soundnails. The bold words are the actual sound source, location, or action. The confusion for sound sources may come from similar materials (e.g. bottle and jar) or textures (e.g. snow and gravel), and similar functions or related interactions (e.g. typewriter and computer). The confusions appear in sound locations can be caused by similar content or participants (e.g. farm and zoo), and similar events (e.g. playground and gym). The confusions for sound-producing actions can result from similar objects involved (e.g. knock and kick) and similar effects they lead to (e.g. crumple and squeeze).

Pre-assigned Concept (bold) and Different Responses	
Source	
alarm: siren, alert, warning, doorbell, clock	**movie**: film, TV, radio, stereo, videogame
baby: infant, newborn, child, kid, toddler, little	**plastic**: wrapper, cellophane, paper
bottle: container, jar, can, dish, plate, glass	**rain**: droplet, storm, hail, waterfall
car: vehicle, engine, truck, bus, motorcycle	**snow**: dirt, leaves, ice, gravel, mud, twig
floor: ground, stairs, porch, patio, surface	**typewriter**: copier,fax,printer,computer
Location	
farm: barn, livestock, ranch, yard, garden, zoo	**school**: class, classroom, college
hospital: clinic,nursery,daycare, medicalcenter	**store**: shop, supermarket, mall, retail
kitchen: restaurant, bar, café, cafeteria, club	**pool**: lake,river,ocean,beach, tub, waterfall
playground: park, court, gym, yard, stadium, theatre, hall, ballpark, auditorium	**train station**: airport, terminal, platform, bus stop, train track
road: street, highway, race track, driveway	**workshop**: factory,garage,construction
Action	
break: crack,creak,crush,shatter,smash,crash	**jingle**: rattle, rustle, fiddle, tinkle, shake
chirp: call, crow, sing, whistle, cackle	**knock**: beat,kick,bang,strike,clap,hit,punch
clink: clank, jingle, tinkle, click, chime	**rub**: scratch,scrub,rip,stretch,twist,squeeze
crumple: scramble, squeeze, rub, crush	**pour**: drip, fill, leak, trickle, splash, drop
crunch: crackle,crisp,rack,scrap,scratch, break	**walk**: gallop,run,jump,stomp,climb,jog, trot
eat: bite, chew, munch, masticate, crunch	**wash**: clean, cleanse, gargle, rinse

Table 6-19. Examples of confusions generated for the sounds

6.3.10 Results and Analysis, Participants

People from 46 countries and regions participated in the sound labeling study. Table 6-20 lists the countries with more than 10 participants. In Table 6-21, the average length of valid tags (removing all function words) and average response times were compared. Significant differences were found in both cases (length of tags: F(8, 1867) = 86.114, $p < 0.01$, $\eta^2 = 0.9885$; response time: F(8, 1867) = 11.833, $p < 0.01$, $\eta^2 = 0.9221$). The results revealed that the response time did not correlate with the length of tags, suggesting that other factors such as proficiency of English may be involved. Even with those differences, responses to the soundnails from people in different countries were similar.

Country	Workers	Country	Workers	Country	Workers
United States	1344	Canada	48	Bahamas	12
India	465	Egypt	24	Philippines	12
United Kingdom	49	Macedonia	15	Germany	11

Table 6-20. Examples of country and worker counts for AMT Soundnail Labeling Study.

Country	Tag Length (words)		Response Time (sec.)	
	Mean	SD	Mean	SD
United States	12.57	1.49	61.07	14.65
India	11.19	1.52	88.30	26.10
United Kingdom	10.86	4.39	64.04	42.24
Canada	12.10	4.00	48.98	26.59
Egypt	18.39	5.62	120.52	144.33

Table 6-21. Valid tag length and response time across countries.

6.3.11 Soundnail Precision Study Discussion

The results from the soundnail labeling study may guide us towards better creation of nonspeech auditory representations.

“If I see the word, I’ll say, of course, it is the sound associated with an umbrella.” The undergraduate students who participated in the pilot sound labeling study stated that given a sound-label pair, the association is often easily established， but given only the sound, retrieving the concept can be difficult. The “Umbrella, Opening” sound is one such example.

It seems while people are concentrating on everyday listening (as expected for our purposes), less information from musical listening is utilized. For example, two soundnails (far away foghorn and far away rooster crowing) were used to illustrate “distance.” The volume of these two sounds is much lower than average, but people still focused on indentifying the source instead of describing the distance. In another example, the “power down” sound is used to evoke “down.” People mostly labeled it as “videogames,” “Sci-Fi,” or “synthesized,” rather than saying the pitch and loudness went down.

The effectiveness of different sounds from similar source(s) may vary greatly. For instance, the top sense score for the “saw – hand saw.wav” soundnail is 1.78, while that for “saw – electric saw.wav” is 0.65; the “train – choochoo.wav” sound (steam train whistling) receives a top sense score of 2.48 while the soundnail “train – arriving.wav” gets a score of 1.41. It implies certain sounds are more distinctive and should be selected as the representation.

Concepts that are ambiguous from an audio characteristics perspective do not seem to have a unique sound associated with them, or at least not a sound distinctive enough at a finer level. For example, a “desk” does not have a characteristic sound of itself, because artifacts do not generate sounds by themselves unless they are deployed by a user; similarly, it seems difficult to distinguish the sound of an iron bell from that of a steel bell, which suggests that fine-grained differences among category members are not audible.

Abstract concepts are hard to evoke. A conceptual-linguistic perspective suggests that many abstract concepts are difficult to evoke with sounds. For example, we tried to represent the concept “day” (meaning a complete 24 hour cycle) by combining a rooster crowing, a clock ticking, and crickets chirping into one sequence. While most participants were able to identify one or more concepts in the sequence, none of them generated the label “day.” Similarly, the sound for “winter” was in most cases labelled “Christmas.” It suggests that for very abstract concepts, people tend to associate the sound with more specific events.

People’s familiarity with the sounds has great impact on their interpretation. This difference may come from several reasons. 1) Age: younger generations have little exposure to old fashioned devices, and thus have more trouble recognize them. For example, the sense score for the “call – rotary dial.wav” soundnail is much lower with the 25 undergraduate students in the pilot study than in the AMT study. 2) Cultural background: people from different cultures may associate completely different sounds with the same event/scene. For example, for labelers from China, the “NBC news theme” which is used to convey the concept “news” may just be a piece of music. 3) Personal experience: people who have never heard an elephant trumpeting are less likely to name the sound correctly.

6.3.12 Soundnail Precision Study Conclusions

In order to verify the efficacy of non-speech auditory representations created for the SoundNet, an online study was carried out using the Amazon Mechanical Turk platform to collect semantic human labels on “what is the source(s) of the sound,” “where are you likely to hear the sound,” and “how the sound is made.” Each of the 327 soundnails was labelled by at least 100 different people. The raw sentences from the study were broken into lexicons, and then stemmed and regrouped by their meanings. Metrics such as sense score, entropy, response time, and length of effective tags were computed to evaluate 1) what concepts can be well conveyed by an environmental sound; 2) what kinds of sounds are distinguishable; and 3) what kinds of confusion may occur in environmental sound perception and what could be the cause.

Results show that, first, sounds that are vocally made by human beings or animals are the most distinctive. They took much shorter time for people to perceive, and there are specific words to describe such as sounds. Comparatively, synthesized, electronic, or machinery sounds are more ambiguous. Second, people can better recognize outdoor scenes than indoor scenes. To be more specific, people, things, and events involved in a school setting or natural environment are more recognisable. Third, sounds made by interactions of two objects are the hardest for people to identify, even worse than sounds with complex scenes and multiple objects for which people simply paid attention to the most distinguish sound source(s).

Overall, 90% of the soundnails succeeded in conveying a concept, among which a third of the agreed concepts were the intended ones included in the non-speech audio-lexicon database. This means that about half of the words in the database vocabulary were verified as audioable. In general, the more concrete the concept is, the more likely a sound can be found to represent it. Concepts related to interactions are more audioable than concepts about locations.

6.4 Evaluation of Non-speech Auditory Representation, a Soundnail Comprehension in Context Study

6.4.1 Research Questions

The sound labeling study described in *Section 6.3* showed that 89% of the soundnails created for SoundNet can convey a concept, and a third evoked the intended concepts. The question arose as to how effective these environmental soundnails are when used to communicate information in a context of common phrases.

A second study "Sounds as Carriers for Communication" was designed and conducted to explore answers to the following questions.

- First, will context improve the performance of soundnails? In the sound labeling study, 46% of the soundnails evoked concepts that were directly related to the sound scenes but differed from those pre-assigned. It is possible that clues from the communication context such as parts of speech could direct people's attention to the intended meaning.
- Second, how well do auditory representations perform compared to pictorial representations? Pictures have long been used in assistive technologies. If we want to apply the data in SoundNet to systems that support communication, we need to verify their effectiveness compared to the use of icons.

Some general hypotheses were proposed before the study.

H1. Visual representations in general outperform auditory representations. As seen in the sound labeling study, many highly imageable concepts are hard to convey via a sound, e.g. "umbrella."

H2. For concepts that are closely related to sound-producing objects/events, audio is a better conveyer than an icon/animation.

H3. People's perception of environmental sounds is improved with the help of sentence context.

H4. The interpretation performance on various parts of speech is less different, since the surrounding words in the sentence have already given some clue.

6.4.2 Study Design

The goal of the *Sounds as Carriers for Communication* study was to investigate people's comprehension of information expressed by non-speech audio in the context of communication. The intent was to investigate how well people could interpret sentences in which words are replaced by soundnails based on SoundNet's audio-concept associations. It merely aims to explore how sounds can convey certain concepts when compared to icons and/or animations. The work described here constitutes constructing and testing a new audio lexicon. Thus it is a proof of concept, not a user study for a specific population.

In the soundnail comprehension in context study, two other stimuli were introduced for comparison. One stimulus used icons (for nouns and adjectives) and the other used animations (for verbs) from Lingraphica. Those iconic representations have been used for almost 20 years in assistive devices to help people with aphasia to compose phrases for language rehabilitation, and therefore, are valid for comparison. In addition, a baseline stimulus which shows a gap in place of the target word tested how much information the context provided. Figure 6-27 shows the example phrase "It is written in the book." in the three different modes.

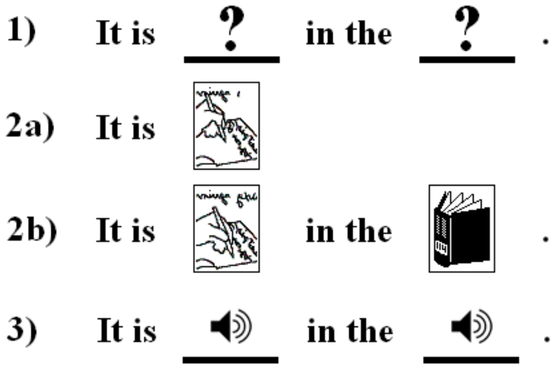


Figure 6-27. Phrases with words replaced by 1) blank, 2a) and 2b) animation, 2b) icon, and 3) soundnail.

Unlike pictures, which can appear at the same time, sounds in a phrase need to be played in sequence. To ensure the proper order, all of the phrases were turned into Flash files, which displayed the words one after another. The interval was one second for context words, and five seconds (the length of the soundnails and animations) for the words replaced with one of the modes. This helped to estimate how much time people spent on interpreting the missing words.

6.4.3 Data Preparation and Study Setting

Eighty-seven concepts in the SoundNet vocabulary with different evocative ratings based on the results from the sound labeling study were selected (Table 6-21). They covered all the cases of soundnail perception listed in Table 6-5 in *Section 6.3.7.1 Pre-assigned Concept vs. Most Agreed-on Concept*.

Perception Accuracy Rating	Number	Example
1	29	buy, day, hair, smoke, travel, etc.
2	27	boat, chalk, fast, rain, walk, etc.
3	24	alarm, bird, cough, ice, print, etc.
4	7	baby, cat, dog, horn, phone, etc.

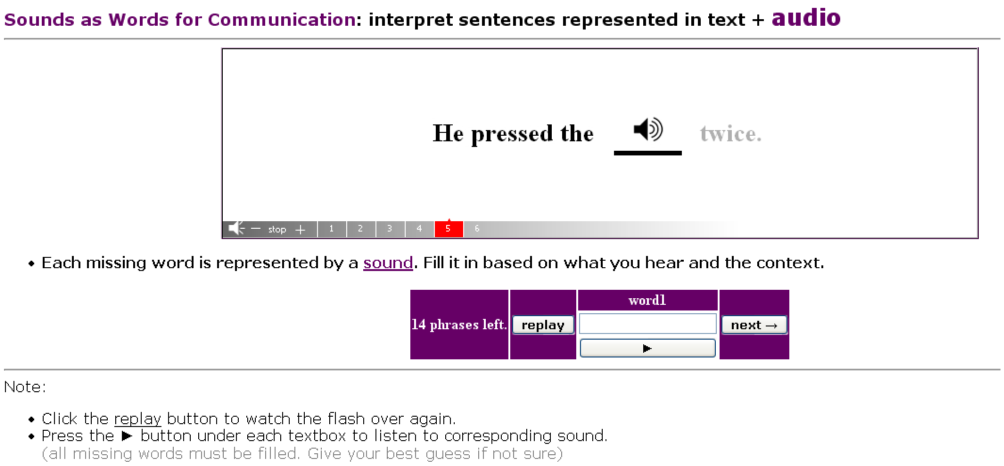
Table 6-21. Selection of target words at different accuracy level in the sound labeling study.

The phrases used in the study came from the Ageless Project. Ageless Project is a blog forum for senior people who fall into the same age span as our ultimate target population, people with aphasia. The posts in Ageless project reflect popular topics among the elderly, and thus is a good reflection of the topics important to the aphasic population and their everyday communication needs. It has been used in other studies described in *Section 4.3* and *Section 5.2* that evaluate comprehension of static and dynamic visual representations in context as well. Sentences with the selected concepts were collected from the blogs. Thirty-six phrases were picked and paraphrased if they were too long (explanations about paraphrasing see *Section 4.3.2.2*). Each phrase was of the length five to twelve words, and had one to four target words embedded.

The study was conducted on Amazon Mechanical Turk. The 36 phrases were divided evenly into nine blocks, and regrouped into 27 HITs. Each HIT contained one block in audio mode, one in icon/animation mode, and one in blank mode. The number of tested concepts within each HIT is relatively even. The mode assignment and position were determined using a Latin Square block design. On the interface (Figure 6-28), the Flash file of a phrase was automatically played. Text fields corresponding to the number of missing words were provided. People were asked to fill in their interpretation of the picture, sound, or gap. They could replay the Flash, as well as individual soundnails in the audio mode. Quality control similar to the sound labeling study was applied (see *Section 6.3.3*). The captcha of someone speaking a sequence of letters and numbers was also implemented in Flash to ensure that participants had proper software installed to play the Flash files. All of the soundnails were converted to Flash, so that people did not need an extra player for the audio files.

	Sentence	Tested Concepts
1	I walk to the gym after I come back from work everyday.	walk,gym,come,work
2	I wear short hair now.	hair
3	It is written in the book.	write,book
4	He sawed the leg off the chair.	saw,chair
5	Do not use a microwave to cook eggs.	microwave,egg
6	It took a day to have the refrigerator fixed.	day,refrigerator
7	I took the train and travelled an hour to get there.	train,travel,hour
8	We saw many turkeys on the farm.	turkey,farm
9	I will bring an umbrella in case it rains.	umbrella,rain
10	Lying in my bed I could hear the wind blowing fast outside.	wind,blow,fast
11	I zipped up my jacket because it was too cold.	zip,cold
12	Festivities include wine tasting, boat cruising, sea plane, and horse riding.	wine,boat,plane,horse
13	I made a phone call to see if the dog was fine.	phone,dog
14	He failed to catch the ball and the window was broken.	catch,ball,window,break
15	She prefers a broom to a vacuum cleaner.	broom,vacuum
16	The bus is available around the clock.	bus,clock
17	I dragged my suitcase across the lobby.	drag,suitcase,lobby
18	I bought a bell for my cat.	buy,bell,cat
19	I brought some balloons to the party for the children.	balloon,child
20	I rewound the movie several times	rewind
21	I heard some horn sound outside my door at night.	horn,door,night
22	He was taken to the emergency room for a surgery.	emergency,surgery
23	The referee whistled every 5 seconds during the basketball game.	whistle,basketball
24	The fire alarm went on while I was cleaning the house.	fire,alarm,clean
25	Using a computer and using a hammer both involve the hands.	computer,hammer,hand
26	Cough associated with smoking does not go away.	cough,smoke
27	Her baby cries a lot, but will laugh when tickled.	baby,cry,laugh
28	It snows a lot here in winter.	snow,winter
29	Obesity can be reversed by eating less and exercising more.	reverse,eat,exercise
30	You can print the picture on paper.	print,paper
31	Birds make sounds similar to music.	bird,music
32	Seatbelts are helpful when the car brakes suddenly.	seatbelt,car,brake
33	I filled my glass with ice and water.	glass,ice,water
34	I am too full after having so many crackers.	full,cracker
35	People can hear my heart pounding from a distance.	people,heart,distance
36	We have run out of chalks.	chalk

Table 6-22. Sentences and concepts tested in the *Sounds as Carriers for Communication* Study.



Sounds as Words for Communication: interpret sentences represented in text + audio

He pressed the ______ twice.

- Each missing word is represented by a sound. Fill it in based on what you hear and the context.

Note:

- Click the replay button to watch the flash over again.
- Press the ► button under each textbox to listen to corresponding sound.

(all missing words must be filled. Give your best guess if not sure)

Figure 6-28. Sounds as Words for Communication interface with a demo sentence.

About 240 people participated in the Sounds as Carriers for Communication study (149 people from the United States, 62 people from India, and the rest from the other 16 countries). Each phrase in each mode was interpreted by at least 50 (up to 74) participants. Effects in different representation modes at both word and phrase levels were tested and compared.

6.4.4 Evaluation Metrics

All typed responses were collected, stemmed, and corrected for misspelling. To better assess the data quantitatively, four evaluation metrics were used. A test for homogeneity of variances in the four metrics showed that results in different modes came from the same normal distribution.

Accuracy rate: the percentage of responses matching the target word, including exact matches and words from the same synonym set (e.g. child, kid).

Response count: the number of different word forms generated by all participants for each sound.

Entropy: the distribution of percentage of word count on different responses. This measures how well people's responses converged. Entropy gives low scores if users agree on a concept and high scores for distributions that are more spread out, which means more words were generated and each has a lower count across all labelers. This takes into account both the total number of different labels (sense sets) that were generated as well as the sense score for each label. Entropy for each sound was computed using the standard equation below, in which p_i was the sense score for label i:

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

Score: the average score of all responses based on the scale in Table 6-23. This includes not only synonyms but also words that are similar and meaningful in the context.

Score	Justification	Example
0	Completely unrelated response.	wood
1	Word in hypernyms, hyponyms, or sister terms in WordNet	yacht
2	Word from the same synonym set.	**boat (intended)**

Table 6-23. Scoring scales and justifications (intended concept “boat”)

6.4.5 Results and Analysis, Word-level Performance

6.4.5.1 Representation Stimuli and Parts of Speech

Test of homogeneity of variances result showed that the performance of the three stimuli came from the same normal distribution. Figure 6-29 shows the number of different responses (F(2,252) = 117.242, $p < 0.001$, $\eta^2 = 0.470$), accuracy rate (F(2,252) = 92.3268, $p < 0.01$, $\eta^2 = 0.415$), entropy (F(2,252) = 107.321, $p < 0.001$, $\eta^2 = 0.447$), and score (F(2,258)=110.500, $p < 0.001$, $\eta^2 = 0.461$) of audio, icon/animation, and blank mode. In all respects, icon/animation stimulus performed significantly better. Entropy difference related to parts of speech of target words is significant (F(2.252) = 3.705, p = 0.026, $\eta_p^2 = 0.788$, $\eta^2 = 0.033$), with responses for noun and verb concepts showing higher convergence than those for adjectives. The small eta squared effect size showed that part of speech was not as great a factor as representation mode. No significant effect was shown in the interaction between representation stimuli and parts of speech, suggesting that given sentence context, parts of speech were not as influential.

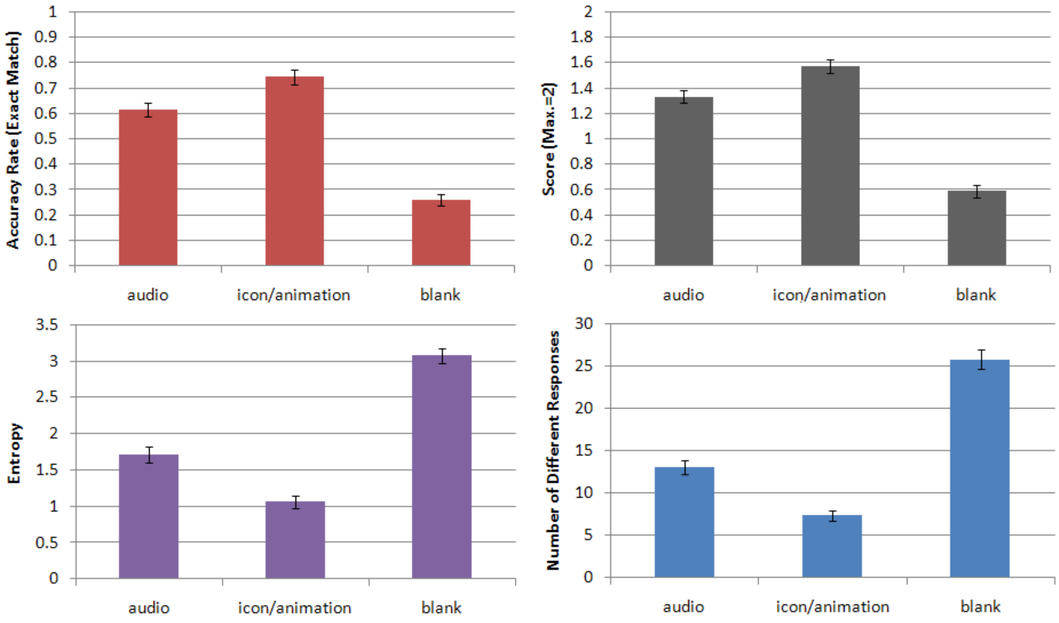


Figure 6-29. Comparison of accuracy rate, score, entropy, and response count across three

Looking at the details more closely and taking entropy as an example, the results for the words can be divided into groups based on the stimulus with the best performance (Figure 6-30). Within the group where the audio mode had lower entropy value (23 concepts), the audio stimulus performed significantly better than the icon/animation stimulus ($F(1,30) = 4.641$, $p = 0.040$, $\eta^2 = 0.809$). Specifically, the audio stimulus significantly outperformed icon/animation stimulus for eight concepts (Figure 6-31) in terms of score, and the scores for another 31 concepts were not significantly different, indicating that certain concepts can be better conveyed by a sound than by an icon or animation.

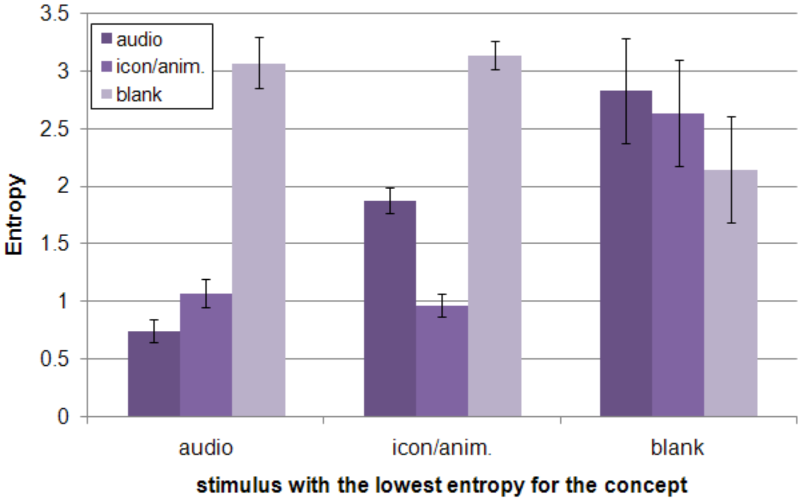


Figure 6-30. Comparison of entropy in different modes within groups categorized by which mode had the lowest value.

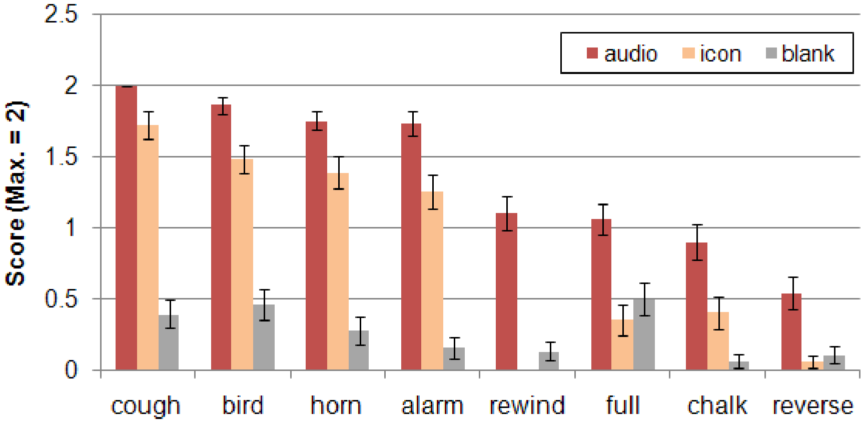


Figure 6-31. Concepts for which the audio score was significantly higher than the icon/animation score.

Among the eight concepts where sounds outperformed icons/animations, two are sound-related artifacts (“horn” and “alarm”), two are things that can generate unique sounds (“bird” with bird songs and “chalk” with the squeaky sound), two are actions that produce distinctive sounds (“cough” and “rewind”), and two are concepts that are commonly associated with highly recognizable sounds (“full” with the burping sound and “reverse” with the truck backing up sound).

Overall, 37 out of 87 auditory representations, 59 out of 87 of visual representations received a score greater than or equal to 1.5, whereas six concepts ("eat," "heart," "night," "rain," "umbrella," and "wine") achieved high scores in the baseline stimulus just by guessing according to the context, for example "wine tasting." There were six iconic representations ("rewind," "reverse," "gym," "full," "chalk," and "lobby," in score-increasing order) with a score of 0.5 and below. The five soundnails with such low score are "gym," "travel," "refrigerator," "suitcase," and "lobby" (in score-increasing order). Further discussion is given in *Section 6.4.7 Soundnail Comprehension in Context Study Discussion*.

6.4.5.2 Influence of Sentence Context

To determine the impact of sentence context on the comprehension of non-speech auditory representations, the results of the audio stimuli from the Sounds as Carriers for Communication study were compared to that from the sound labeling study. As shown in Figure 6-32, the target words with high sense scores in the previous study were again those with significantly higher accuracy rate than the ambiguous ones (F(1,85) = 37.037, $p < 0.001$, $\eta^2 = 0.303$).

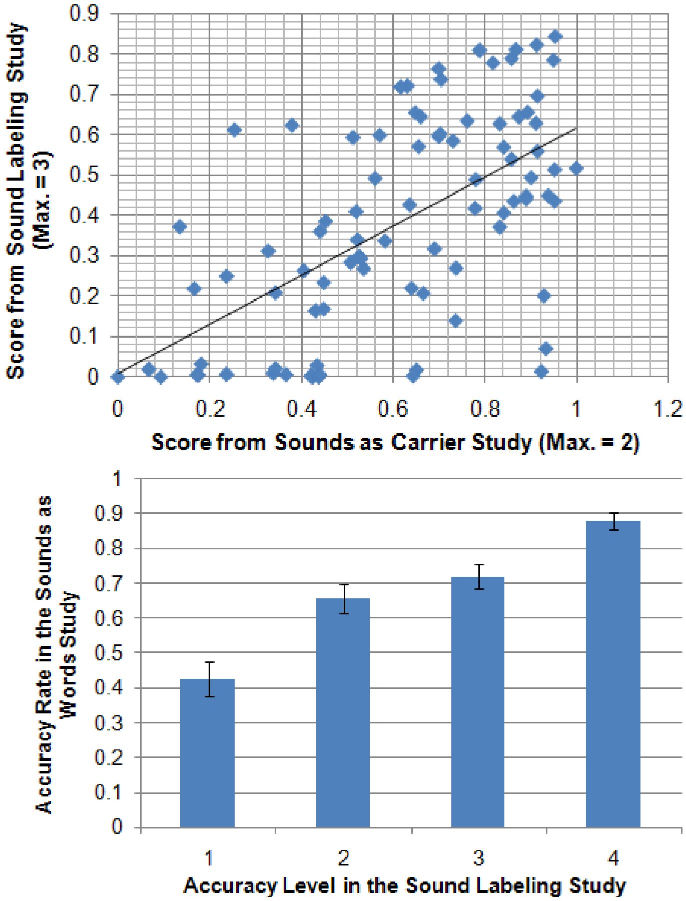


Figure 6-32. Comparison of soundnail perception accuracy with and without context.

However, context did provide information for people to identify the sounds or concentrate on intended aspects in the auditory event in many cases. Table 6-24 lists the 10 words with highest accuracy rate in audio mode as well as their corresponding blank mode accuracy rate. Six out of the ten words had an accuracy level of 1 or 2 in the sound labeling study, and half of them (particularly those at accuracy level 1) had an accuracy rate higher than 0.7 in the blank mode. This meant that people could guess these words quite well purely based on the context. An example is "I will bring an umbrella in case it rains." In other cases, the context suggested the part of speech of the missing word. For example, the "baby crying" sound was used to illustrate the word "cry." In the sound labeling study, many people identified the sound as "baby." The phrase given in the second study was "Her baby ____ a lot …" which indicated that the missing word should be a verb. As a result, people mostly generated "cry" instead of "baby."

Words	Audio Accuracy (with context)	Blank Accuracy (with context)	Labeling Accuracy (without context)
cough	1.0000	0.1970	3
cat	0.9545	0.1167	4
cry	0.9531	0.3788	2
laugh	0.9531	0.4242	2
dog	0.9508	0.2500	4
rain	0.9394	0.8919	2
wine	0.9342	0.8000	1
night	0.9298	0.7200	1
umbrella	0.9242	0.9054	1
eat	0.9153	0.8571	3

Table 6-24. Comparison of accuracy in audio and blank stimuli for the top 10 words with highest auditory accuracy

6.4.5.3 Linguistic Properties: Concreteness and Imageability

Concreteness and imageability had significant impact on the perception and interpretation of soundnails, however, the effect was less strong compared to the soundnail precision study result (Table 6-25). With the help of context, the accuracy rate of abstract words was greatly increased (Figure 6-33 left). The accuracy rate of words with an average level of concreteness (=0) even approached highly concrete ones. Similar effect was found in imageability (Figure 6-33 right).

	concreteness			imageability		
	F(1, 85)	p-value	η^2	F(1, 85)	p-value	η^2
without context	13.821	< 0.001	0.140	10.085	0.002	0.106
with context	4.920	0.029	0.055	8.804	0.006	0.087

Table 6-25. Comparison ANOVA results on concreteness and imageabilty with and without context.

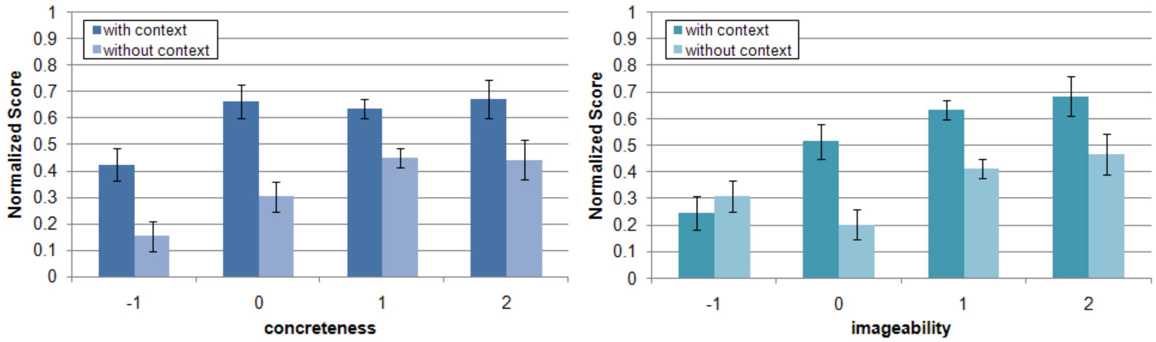


Figure 6-33. Concepts for which the audio score was significantly higher than the icon/animation

6.4.6 Results and Analysis, Sentence-level Performance

Phrase level results were similar to the word level. The average score of target words in each sentence was computed, and the icon/animation mode significantly outperformed the audio mode (F(2, 105) = 62.493, $p < 0.001$, , $\eta^2 = 0.543$, Figure 6-34 left). However, there were still five sentences for which the scores in the audio mode was significantly higher than the visual mode (Figure 6-35, Table 6-26). The concepts that are identified to be more audioable than imageable appear in the five sentences (“horn,” “alarm,” “rewind,” “chalk,” and “full”). The correlation between number of tested concepts in the sentence and average score is not that strong (0.138 for audio and 0.266 for icon/animation).

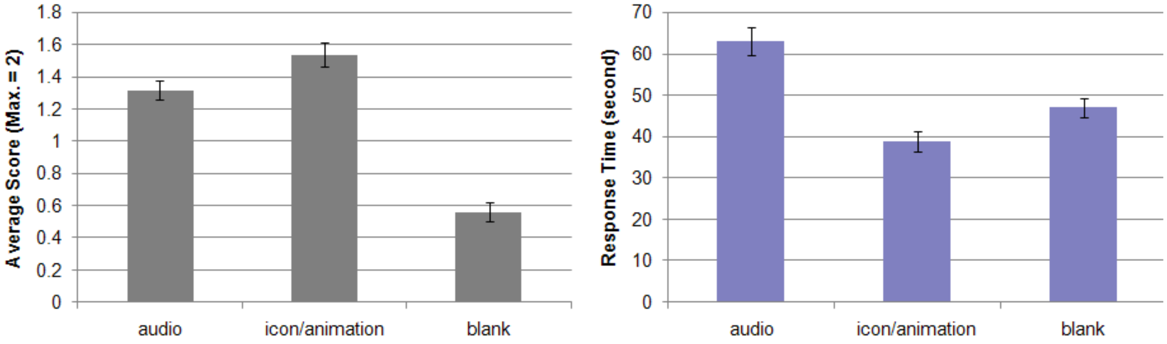


Figure 6-34. Comparison of phrase level score and response time across three stimuli.

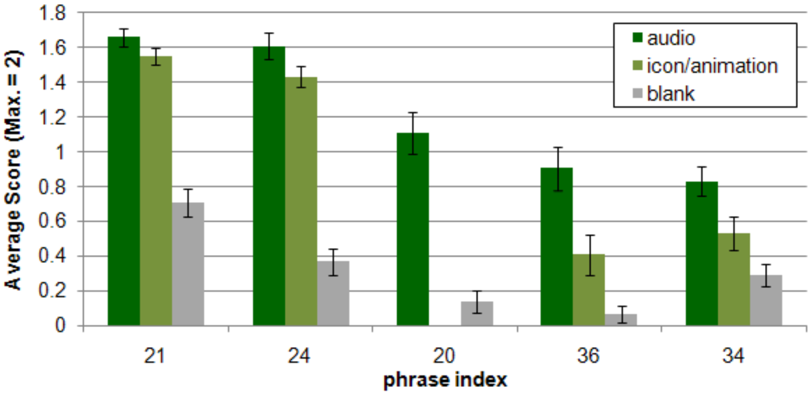
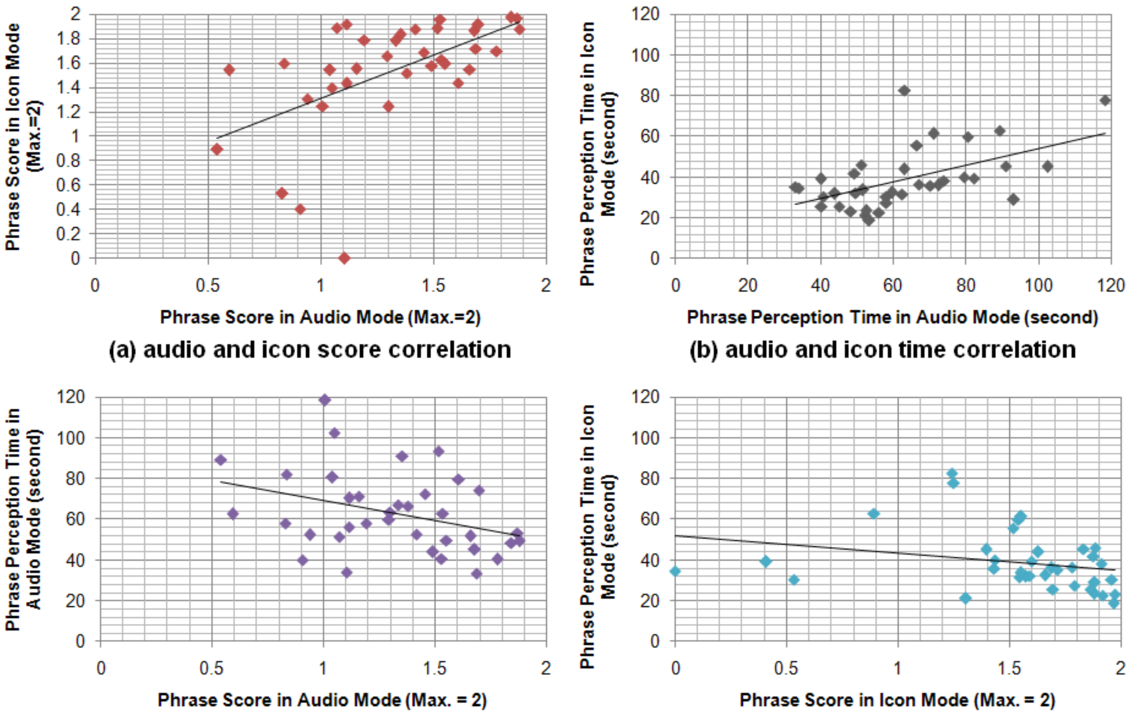


Figure 6-35. Phrases for which the audio score was significantly higher than the picture score.

21	I heard some horn sound outside my door at night.
24	The fire alarm went on while I was cleaning the house.
20	I rewound the movie several times.
36	We have run out of chalks.
34	I am too full after having so many crackers.

Table 6-26. Phrases for which the audio score was significantly higher than the picture score.

The response time was computed for each phrase, calculated as the time between phrase loading and the response submission (logged by the interface) minus the time spent on playing sounds for context words. Although the response time could be affected by participants' behavior in the study (for instance, some started to type as soon as the sound began to play, while others waited until the sound finished playing), it still provides a rough estimate of how long people spent on trying to figure out the missing words and typing in the answers. Figure 6-34 (right) showed that overall, significantly more time was required for the audio mode ($F(2,105) = 20.279$, $p < 0.001$, $\eta^2 = 0.279$). Furthermore, the correlation between number of tested concepts in the sentence and the response time in audio stimulus is 0.71, whereas the correlation is 0.58 for the iconic stimulus. It suggests that unlike pictures, which people can interpret at a glance, sounds may require listening to the entire clip before forming an idea.



(a) audio and icon score correlation

(b) audio and icon time correlation

(c) audio score and time correlation

(d) icon score and time correlation

Figure 6-36. Correlation between phrase score and response time with the audio/icon stimuli.

In general, a phrase that is hard to interpret and takes longer time given the iconic representations is difficulty for sound representations as well (Figure 6-36(a)(b)). However, in the audio mode, time spent on words for which people showed low agreement was not significantly longer than that spent on words where people showed high agreement (Figure 6-36(c)). Similar situation happed with icons/animations as well, though the average response time is much shorter (Figure 6-36(d)). This suggests that time might be an important feature for auditory representations, whether the sound was recognizable or not.

6.4.7 Discussion

There were a few interesting facts observed in the studies.

An essential question is how to illustrate abstract concept with sounds. Even with surrounding context, it is still challenging for listeners to conclude an abstract concept from an environmental sound. When trying to evoke the word "day" with a sound playing rooster crewing, clock ticking, and crickets chirping in sequence, most people put down "rooster" even though the phrase was "It took a day to have the refrigerator fixed."

As mentioned earlier, there were six icons/animations ("rewind," "reverse," "gym," "full," "chalk," and "lobby"), and five soundnails ("gym," "travel," "refrigerator," "suitcase," and "lobby") with a score of 0.5 and below. The common ones "gym" and "lobby" are both places, which may suggest that location in general is hard to depict both visually and with sounds. Two concepts, "refrigerator" and "suitcase," which sounds fail to convey, are both highly imageable and commonly encountered in everyday life. It suggests that even if a sound event is familiar to the listeners, as long as its acoustic characteristics is not that distinctive, people have trouble recognizing it without any visual clue. What makes things worse is, "suitcase" and "lobby" as well as "refrigerator" and "day" (another poorly audioable concept) appeared in the same sentence. These concepts are hard to recognize given a sound by themselves to begin with. Adding another confusing sound can only increase the difficulty in interpretation.

Table 6-27 and 6-28 listed some responses generated for sample iconic/auditory representations. Two kinds of answers were generated when confusion occurs: guesses based on the visual/auditory content or induction from the context ignoring the given representation. If the content itself is clear (regardless of the pre-assigned concept), people are more likely to come up with answers in the first category. For example, almost every participant recognized the fire icon for the phrase "fire alarm," still, many of them considered that the alarm clock icon has something to do with time, even if it does not make sense in the context. On the contrary, if the context is more straightforward than the representation, people may try to piece together information using the surrounding word, regardless of what picture/sound is shown/played. An example sentence is "Obesity can be reversed by eating less and exercising more."

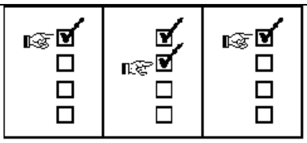
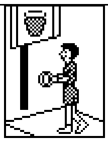
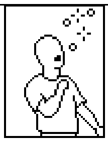
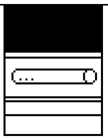
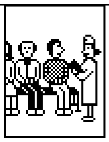
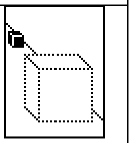
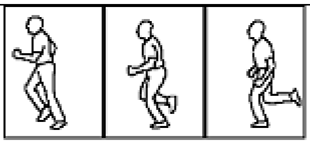
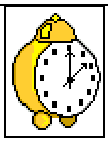
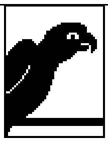
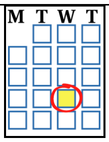
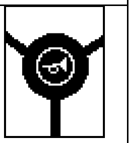
	rewind	gym	full	chalk	lobby	distance
Pic.						
	rewind, check watch, see select, choose pick, buy, vote screen, review	gym basketball court park playground hoop stadium	full, sick bath,cold heavy weak sleepy lazy thirsty	chalk cigarette cigar classroom smoke lock stick	lobby,room crowd, line hospital office,clinic bench people patient, floor	distance box corner block device far mile
	reverse	**alarm**	**bird**	**cough**	**day**	**horn**
Pic.						
	reverse, reduce, control cure, stop, help, beat run, jog, walk, pace, treat, outpace, defeat prevent, skip, overcome conquer, manage	alarm time clock ring timer 2 o'clock	bird parrot quail blackbird raven crow	cough breathe chew asthma sickness cancer	day month week appointment calendar Wednesday	horn loud trumpet music beep weird

Table 6-27. Examples of responses to the confusing icons/animations (Pic).

	rewind	gym	full	chalk	lobby	distance
Aud.	rewind, skip tape	gym, workout breathing	full, burping	chalk, write on board	lobby, airport lobby announcement	distance, rooster call in distance
	rewind watch record forward unzip play	gym, park store, car lift, heart breathe train, stairs factory	full belch, gassy burp, bloat stuff, load fill, sick, rude erupt, yawn	chalk,shave, scraper pencil, tape fire, Kleenex turkey, tissue wood, water	lobby, hallway airport, station terminal, floor hall, room, road bridge, office people, city	distance rooster, crow morning, far corner, farm foot, shoe away, distant
	reverse	**alarm**	**bird**	**suitcase**	**day**	**refrigerator**
Aud.	reverse, truck back up	alarm, fire alarm	bird, nightingale sing	suitcase, clip open	rooster crowing clock ticking crickets chirping	refrigerator, close door bottle clinking
	reverse reduce stop back up bell danger	alarm buzzer ambulance siren bell worry	bird hummingbird quail parrot	suitcase, belt stapler, type luggage, can purse, lighter, guitar, rock car, shotgun	day rooster, hen time, while morning, hour horse, alarm wake up	refrigerator car, door, food machine, oven lunch, kitchen scramble,egg pump,sink,

Table 6-28. Examples of responses to the confusing sounds (Aud.).

People’s interpretation of both pictorial and auditory representations is more or less affected by the prototype phenomenon. Responses to the concept “bird” are a good example. Even if the bird icon is quite generic, people still managed to catch little details like the beak and the color to draw more specific conclusions. A similar but less severe situation happened with the nightingale sound for the “bird” concept. It was because most of the listeners were not able to recognize the particular species of the bird and thus could only conclude that it is a bird. A seagull sound may not work well in the “bird” case since the labeling study showed that people can identify the unique squawking sound as from a gull. It suggests that a too distinctive sound from its subordinates may not be as good as a representation for a more generic concept, although most of the time people cannot make such fine discrimination (e.g. iron and steel, ceramic and china).

6.4.8 Conclusions and Guidelines for Non-speech Auditory Representation Design

The soundnail labeling study introduced in *Section 6.3* collected semantic human labels on the source(s), location(s), and interaction(s) of 327 soundnails. A further study “Sounds as Carriers for Communication” was conducted to evaluate the efficacy of environmental sound representations in daily phrase context in comparison to icons and animations. Results showed that although the icon/animation mode had better performance overall (hypothesis H1 confirmed), there were seven concepts for which the audio mode had significantly higher scores, while there were another 31 words for which the auditory and visual modes were not significantly different (hypothesis H2 confirmed). This suggests that audio has advantages in conveying certain concepts over pictures and may be utilized in assistive systems. Context did improve people’s perception of soundnails, even for abstract concepts (H3 confirmed). Part of speech is no longer a factor with significant influence given the clues embedded in surrounding words (H4 confirmed).

Based on analysis and observation from the two studies, I propose the following guidelines for designing effective non-speech auditory representations for frequently used concepts:

1. Avoid using synthetic sound if possible. People almost always associate them with “science fiction,” “movie,” “video game,” or “aliens.”
2. Minimize the use of acoustic properties instead of sound events to convey information. Without notice, listeners might not notice that the change of pitch for instance indicates direction. It may require some learning as well.
3. Use familiar sounds, and be sensitive to age and cultural differences.
4. Utilize distinctive acoustic characteristics if available. Choose a more unique sound for the same concept, e.g. the handsaw is better than an electric saw for “saw” as a tool or an action.
5. Be careful of the use of the unique sound of a sub category to depict a more generic concept.

CHAPTER 7 MULTIMEDIA LANGUAGE APPLICATIONS

Chapter 4, Chapter 5, and Chapter 6 have provided the theories behind the viable use of visual and non-speech auditory representations as concept carriers. Along the way, a multimedia augmented semantic network, which consists of over 5,000 web images, over 3,000 icons, about 70 animations, 50 videos, and over 320 environmental sounds that are associated with frequently used concepts (in the form of content words), was constructed. As the next step, real-life scenarios are introduced to test the efficacy of communication support via this multimedia language.

This chapter first introduces the web-based communication support system (called Online Multimedia Language Assistant) that serves as the interface to the multimedia language. Then evaluations of the system in two usage scenarios, understanding of online information about food and medical care by non-native English speakers, are presented. General findings are discussed with a number of system modification proposed. The main goal of this work is to facilitate web information comprehension for people with language barriers or disabilities. As a side product, given the easy access to the multimedia library, people encountering communication difficulties can use any available visual/auditory representation to assist on-going conversations.

7.1 Online Multimedia Language Assistant (OMLA)

The Online Multimedia Language Assistant (OMLA) for individuals with language disabilities or barriers provides a platform accessible over the Internet for applications such as looking up unknown words, constructing phrases for communication, practicing pronunciations, and accessing content. OMLA also enables resource sharing and remote collaboration.

7.1.1 Conventional Approaches to Resolving Communication Barriers and Their Limitations

Chapter 2 has pointed out the verbal communication difficulties that people with impaired/non-proficient language skills face everyday. There is another kind of communication difficulty that comes from information exchange through text. For instance, the Internet is now an essential source of information; however, there is an imbalance between the distribution of Internet language and the user population: over half of the webpages are written in English, while only about 30% of Internet users speak English as their first language (internet statistics2009). For people whose native language is different from the language in which the webpages are written, the effectiveness of comprehending information diminishes as unknown words occur. Such users encounter all sorts of problems and frustration when browsing the web or using online services such as online shopping or news.

When an unfamiliar word that hinders the understanding of the whole message occurs or when the word in the non-native language is unknown, often people have to look it up in paper or online dictionaries (e.g. Babelfish or Google Translator). However, a dictionary or translation does not always help. First, there could be no translation available for what one wants to look up. For example, no results were returned when I searched for “define: vegetarian chicken” on Google (in March 2010). Second, often the translated word makes no sense (e.g. in Figure 7-1, the Chinese word for “water chestnut” gets translated into “horseshoe)? Often translation dictionaries are of little use because knowledge of the domain is also required. What make things worse is, sometimes people do not even know the term in their own language. For example, many people in the south of China do not know the other name of water chestnut in Chinese (the one that is translated in to “eleocharis” in Figure 7-1). Also, if more unknown words appear in the definition, people may get even more confused. In most cases, a human is better at identifying the meaning from the context than a book or a machine, however, confusion can occur. For example, if I tell you that I am looking for the name of an edible tuber that is dark brown outside and white inside, what would you think of? Many people might say “potato,” especially if they are not familiar with water chestnut. Such complicated information processing is challenging even for a person with unimpaired language skills, and it becomes impossible for individuals with language disabilities or language barriers.

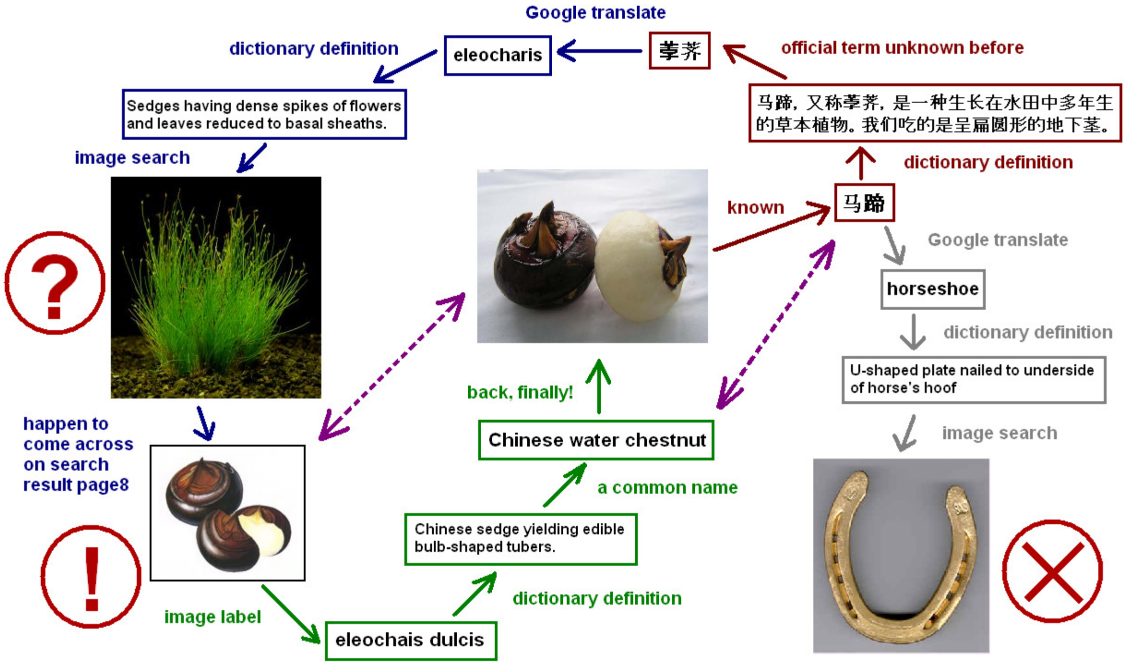

Google translate
dictionary definition
eleocharis
荸荠
official term unknown before
Sedges having dense spikes of flowers and leaves reduced to basal sheaths.
马蹄，又称荸荠，是一种生长在水田中多年生的草本植物。我们吃的是呈扁圆形的地下茎。
image search
dictionary definition
known
马蹄
Google translate
horseshoe
dictionary definition
U-shaped plate nailed to underside of horse's hoof
?
back, finally!
happen to come across on search result page8
Chinese water chestnut
image search
a common name
!
Chinese sedge yielding edible bulb-shaped tubers.
dictionary definition
image label
eleochais dulcis

Figure 7-1. A search marathon.

There are situations where the correct translation and definition is given, but with more unknown words contained in it. This makes people more confused. Even if the person understands every single word given in the definition, he/she may still have no clue to its meaning. Here is an example. I was chatting with a friend online. She was talking about a boxer that she saw on the street the other day. From the context, I sort of figured out that the "boxer" was not someone who fights for sport. Kindly she clarified that it was a dog, but I still could not quite follow the conversation since I did not know what a boxer dog is like. So I first looked it up in the dictionary as the chat went on. "A breed of stocky medium-sized short-haired dog with a brindled coat and square-jawed muzzle developed in Germany" was what I found in WordNet. I got blocked by the words "brindled" and "muzzle," so I turned to the internet for further help. I found a very detailed description on the web (Table 7-1) which I can understand pretty well. However, I still could not piece all the information together to form an image of the dog in my head. Besides, after all these readings, I was completely disconnected from the conversation with my friend. This is definitely a case where a picture is worth a thousand words. All I need was a picture of a boxer (Table 7-1) to help me quickly pick up the idea and move along in the communication.

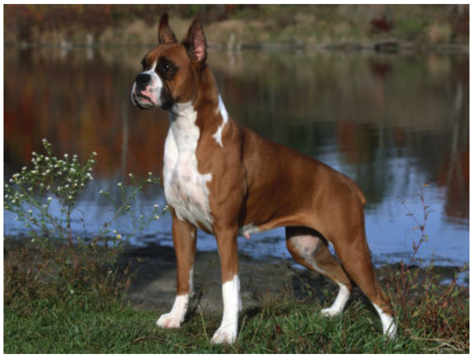
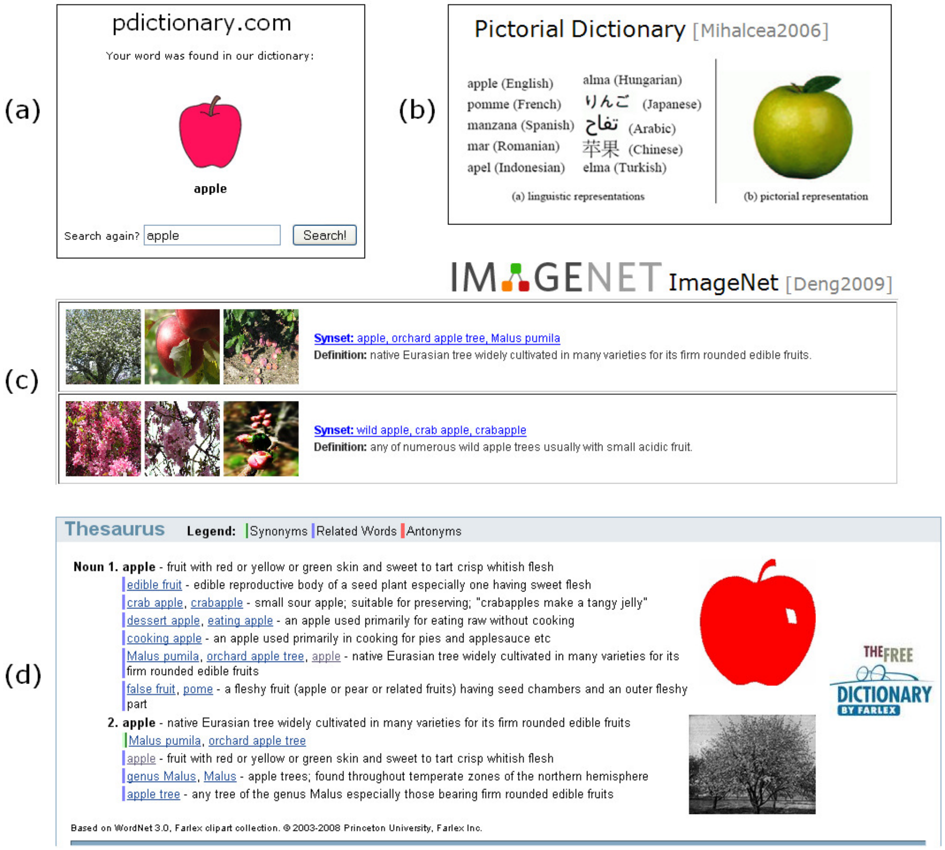
Verbal Description (from dogbreedinfo.com)	Picture
The Boxer's body is compact and powerful. The head is in proportion with the body. The muzzle is short and blunt with a distinct stop. The nose is large and black with very open nostrils. The jaw has an under bite. The eyes are dark brown. The ears are set high, either cropped or kept natural. When left natural the ears are thin, falling forward, laying close to the head. The neck should be round, strong, and muscular, and without dewlap. The muscular, front legs are straight and parallel when viewed from the front. The back legs are well muscled. The tail is set high and usually docked. Dewclaws are usually removed. The short, smooth, close-fitting coat comes in, fawn, brindle, tan, mahogany, black often with white markings.	

Table 7-1. Verbal descriptions and picture of "boxer" from the Internet.

With demand, various picture dictionaries have come into being. The Oxford Picture Dictionary is an example of the paper-based dictionary that uses illustration to help explain terminologies. There are many other computer-based picture dictionaries available for more general user populations in one language (pidic.com2009), and multiple languages (e.g. pdictionary.com2009) (Figure 7-2(b) PicNet Borman2005). In the past, online picture dictionaries assigned pictures to (polysemous) words (Figure 7-2 (a)), similar to the AAC vocabularies described in *Section 2.3*. Nowadays, many online picture dictionaries (freeonlinedictionary.com2010) have evolved to be more sophisticated. They have a larger glossary, with mapping from pictures to meanings (specific definition) (Figure 7-2(d)). However, the majority of the existing picture dictionaries merely present information in a one-picture-to-one-word format, and the data is mainly collected and maintained by the developers. PicNet (Borman2005)(Mihaclea and Leong2006) and ImageNet (Figure 7-2(c), Deng et al.2009) are attempts to extend WordNet with pictures using low-cost human labor available on the Internet. Most of the web pictorial dictionaries are implemented as a website that requires users to open a separate page for the search instead of retrieving the results instantly. As mentioned in the example with “boxer,” such switching back and forth between communication and dictionary context may affect the fluency of the communication.


(a)
pdictionary.com
Your word was found in our dictionary:
apple
Search again? apple
Search!
(b)
Pictorial Dictionary [Mihalcea2006]
apple (English)
pomme (French)
manzana (Spanish)
mar (Romanian)
apel (Indonesian)
alma (Hungarian)
りんご (Japanese)
تفاح (Arabic)
苹果 (Chinese)
elma (Turkish)
(a) linguistic representations
(b) pictorial representation
IMAGENET ImageNet [Deng2009]
(c)
(d)
Thesaurus
Legend: Synonyms Related Words Antonyms
THE FREE DICTIONARY BY FARLEX

Synset: apple, orchard apple tree, Malus pumila
Definition: native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits.

Synset: wild apple, crab apple, crabapple
Definition: any of numerous wild apple trees usually with small acidic fruit.

Noun 1. apple - fruit with red or yellow or green skin and sweet to tart crisp whitish flesh
edible fruit - edible reproductive body of a seed plant especially one having sweet flesh
crab apple, crabapple - small sour apple; suitable for preserving; "crabapples make a tangy jelly"
dessert apple, eating apple - an apple used primarily for eating raw without cooking
cooking apple - an apple used primarily in cooking for pies and applesauce etc
Malus pumila, orchard apple tree, apple - native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits
false fruit, pome - a fleshy fruit (apple or pear or related fruits) having seed chambers and an outer fleshy part

2. apple - native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits
Malus pumila, orchard apple tree
apple - fruit with red or yellow or green skin and sweet to tart crisp whitish flesh
genus Malus, Malus - apple trees; found throughout temperate zones of the northern hemisphere
apple tree - any tree of the genus Malus especially those bearing firm rounded edible fruits

Based on WordNet 3.0, Farlex clipart collection. © 2003-2008 Princeton University, Farlex Inc.

Figure 7-2. Examples of online picture dictionaries, search term “apple.”

The Online Multimedia Language Assistant (OMLA) developed in my thesis research differs from previous work in that it uses the web browser extension as the system platform which can be accessed simultaneously with the webpage being viewed; it is strengthened by a backend multimedia database that maps multiple visual representations to each concept in the vocabulary; and users can submit personal images, videos and sounds for adaptably and adaptively expanding the underlying multimedia language.

Similar strategies have been used in second language learning to improve reading comprehension. For example, a software template (Ariew2006) was design to generate hypermedia texts for the use of foreign and second language students. Reading materials were annotated with hypertext (text, graph, audio, and video) by the materials designer or second language acquisition researcher, and presented to the students. Once a highlighted annotated word was selected, a menu listing the types of available hypertext annotation popped up for the student to choose to view.

OMLA no longer restricts the availability of visual representations to a specific designed reading material for second language students. It aims to facilitate information comprehension and delivery in everyday setting. OMLA also involves the viewers of the dictionary in its vocabulary creation and maintenance.

7.1.2 System Overview

The Online Multimedia Language Assistant (OMLA) consists of three main parts (Figure 7-3): a library of rich multimedia-concept associations consisting of images, icons, animations, videos, and environmental sounds, an adaptable and adaptive vocabulary (ViVA: Visual Vocabulary for Aphasia) which enables efficient vocabulary navigation and word retrieval, and web interfaces for users to navigate and access library items. ViVA is designed by Sonya Nikolova (Nikolova et al. 2010) from the Princeton Aphasia Project. Currently, ViVA is constructed and tested separately, and the OMLA system is at its early stage of integration.

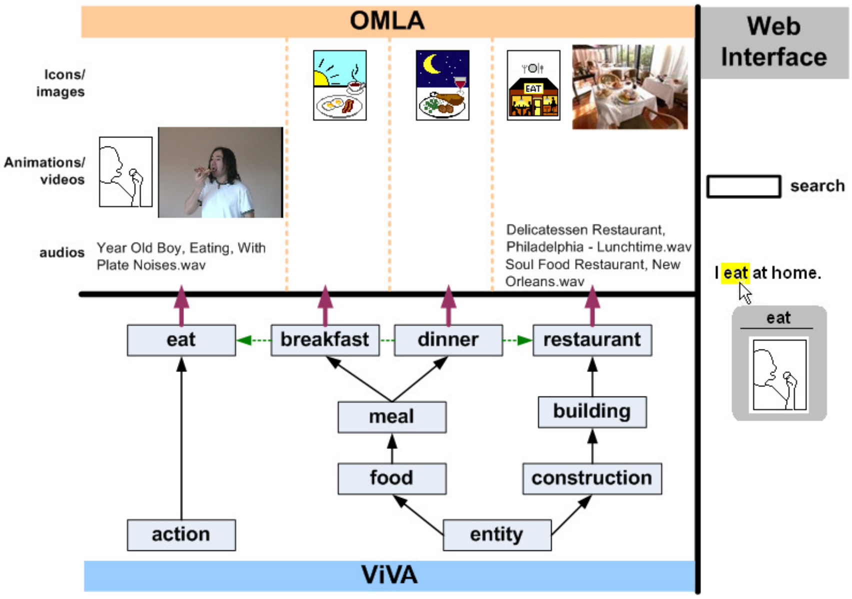
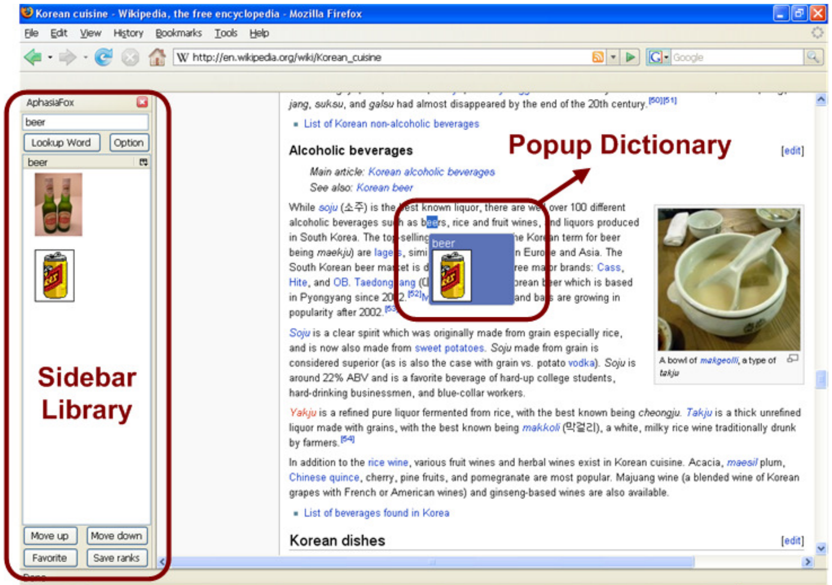


Figure 7-3. OMLA architecture.

The interface of OMLA is designed and implemented as a Mozilla Firefox extension with all the multimedia data managed in a central server. The interface (Figure 7-4) attached to the Firefox window consists of three parts: a popup display of visual representations for selected words, a sidebar view of the backend library, and a dialog window for uploading images. Figure 7-5 shows the overall information and control flow among the different components. Two undergraduate students Zennen Clifton (Princeton'2008) (Clifton2007) and Brett Lullo (Princeton'2011) (Lullo2009) participated in the implementation of the OMLA interface.


Korean cuisine - Wikipedia, the free encyclopedia - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://en.wikipedia.org/wiki/Korean_cuisine
Google
AphasiaFox
beer
Lookup Word
Option
beer
Sidebar Library
Move up
Move down
Favorite
Save ranks
Popup Dictionary
beer

jang, suksu, and galsu had almost disappeared by the end of the 20th century.[50][51]

- List of Korean non-alcoholic beverages

Alcoholic beverages [edit]

Main article: Korean alcoholic beverages
See also: Korean beer

While soju (소주) is the [b]est known liquor, there are we[ll] over 100 different alcoholic beverages su[c]h as beers, rice and fruit wines, [a]nd liquors produced in South Korea. The to[p]-selling ... [t]he Kor[e]an term for beer being maekju) are lage[r]s, simi ... n Euro[p]e and Asia. The South Korean beer ma[r]ket is d ... ree ma[j]or brands: Cass, Hite, and OB. Taedong[g]ang (... [K]orean b[e]er which is based in Pyongyang since 20[0]2.[52] M ... and ba[r]s are growing in popularity after 2002.[53]

Soju is a clear spirit which was originally made from grain especially rice, and is now also made from sweet potatoes. Soju made from grain is considered superior (as is also the case with grain vs. potato vodka). Soju is around 22% ABV and is a favorite beverage of hard-up college students, hard-drinking businessmen, and blue-collar workers.

A bowl of makgeolli, a type of takju

Yakju is a refined pure liquor fermented from rice, with the best known being cheongju. Takju is a thick unrefined liquor made with grains, with the best known being makkoli (막걸리), a white, milky rice wine traditionally drunk by farmers.[54]

In addition to the rice wine, various fruit wines and herbal wines exist in Korean cuisine. Acacia, maesil plum, Chinese quince, cherry, pine fruits, and pomegranate are most popular. Majuang wine (a blended wine of Korean grapes with French or American wines) and ginseng-based wines are also available.

- List of beverages found in Korea

Korean dishes [edit]

Figure 7-4. Web Interface components of the Online Multimedia Language Assistant.

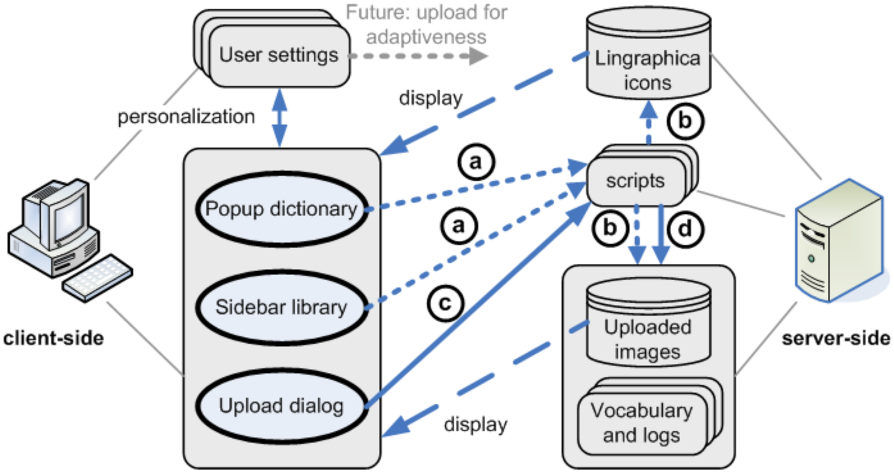


Figure 7-5. Information and control flow: (a) retrieve queried text; (b) look up word in the vocabulary and return associated pictures; (c) upload image from local; (d) create thumbnail, store uploaded picture, and update vocabulary.

OMLA provides a web-based popup visual dictionary that allows web-surfers to access multimedia augmented definitions without going through the trouble of switching between websites or consulting a paper-based picture dictionary. Furthermore, OMLA offers a platform for people to view the multimedia-lexicon library supporting the dictionary in the backend, to customize the display of the dictionary, and to share personal images. The customization process involves users in active multimedia learning (Mayer2001). In addition, their input can later be used as training data for machine learning algorithms to improve the image retrieval process for other users. The following sections introduce each of the OMLA interface components separately.

7.1.3 Popup Multimedia Dictionary

After OMLA is activated through the menu, the popup box (Figure 7-6) displaying a representative picture is triggered when the user picks a word on the webpage. The popup mechanism is based on an open-source Firefox add-on, a Japanese dictionary called Rikaichan (polarcloud.com2008).

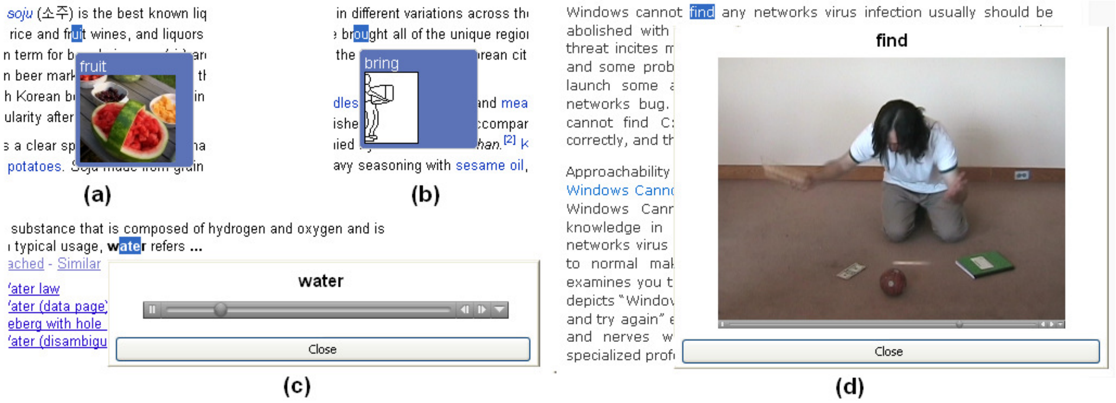


Figure 7-6. Popup dictionary with (a) image, (b) animation, (c) sound, and (d) video.

The initial design displayed pictures on mouse roll-overs, which caused word search results to continuously appear. We modified the program to toggle the popup by slowly clicking on or selecting part of a specified word. OMLA obtains the current mouse location at a mouse down event, and then branches outwards on both sides until reaching white space, punctuations, or special boundary symbols. The extracted string is then sent to the server for further processing. We chose partial selection over exact entire word selection because the latter usually takes more effort to manipulate the mouse.

The search algorithm is based on the NLTK-Lite (NLTK2009) lexicon morphing function, first comparing the retrieved text to the initial vocabulary of over 2000 words compressed from Lingraphica as well as the new words added by the users. If the highlighted word does not match an image in the database, we lemmatize the word using NLTK and search for the lemmatized word. The appropriate list of icons and/or uploaded images is returned to the client if the word is successfully retrieved from the dictionary (Figure 7-6(b) "brought"). A box with the representative picture/sound/video (according to the user media type option) and base form of the queried word pops up near the cursor. For images, icons, and animations, the popup box disappears automatically when the user moves the mouse again. For sounds and videos, the user has to click a button on the box to close the popup. This is designed this way because sound and video files need to be streamed and they require a few seconds to play.

AphasiaFox also allows compound (two to three) words to be recognized from the webpage (Figure 7-7). In this setting, the word crawling window is stretched to include one word before and one word after the word at the current mouse down location. The three words (A, B, C) are sent to the server, and three combinations (AB, BC, ABC) are checked to see if they match any compound word in the dictionary.

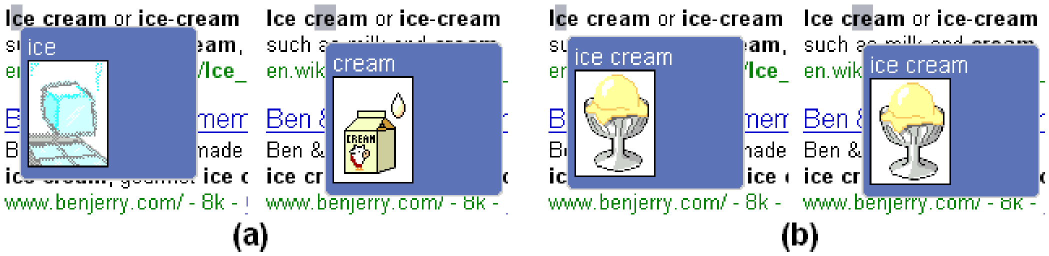
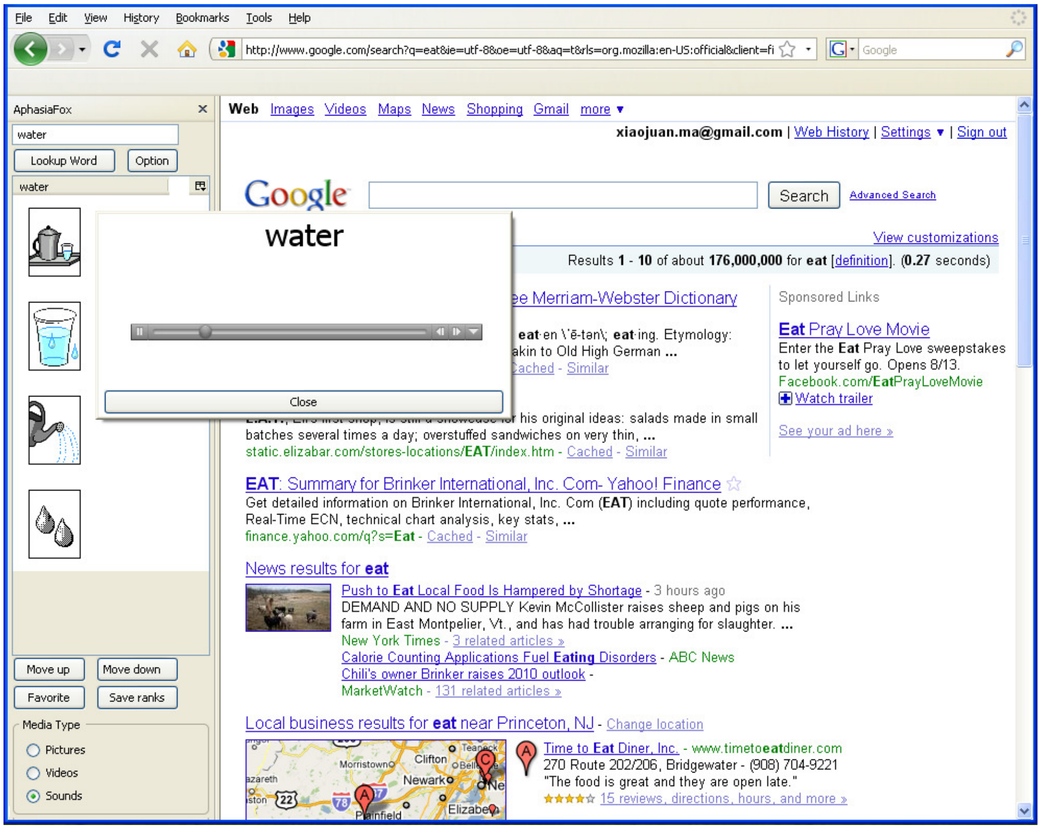


Figure 7-7. Example ("ice cream") of text retrieval modes: (a) individual words; (b) compound

7.1.4 Adaptable Sidebar Multimedia Library

7.1.4.1 Sidebar Library

When a word included in the vocabulary is searched in the sidebar multimedia library, all the associated visual representations appear in a fixed order: uploaded pictures sorted by submit time come before icons sorted by index, followed by a list of environmental soundnails, and video (if any) at last. Which representation shows up in the popup display is based on the media type specified by the user. Picture is the default setting, and is always listed in the sidebar even if the popup display is in audio or video mode (Figure 7-8). We designed it this way reasoning that even if people cannot recognize the sound or video, they still receive pictorial clue from the images/icons/animations.


File Edit View History Bookmarks Tools Help
http://www.google.com/search?q=eat&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=fi
Google
AphasiaFox
water
Lookup Word
Option
water
water
Close
Move up
Move down
Favorite
Save ranks
Media Type
Pictures
Videos
Sounds
Web Images Videos Maps News Shopping Gmail more ▾
xiaojuan.ma@gmail.com | Web History | Settings ▾ | Sign out
Search
Advanced Search
View customizations
Sponsored Links
See your ad here »
Morristown
Clifton
Newark
Elizabeth
Plainfield
Teaneck

○ Pictures
○ Videos
● Sounds

Results **1** - **10** of about **176,000,000** for **eat** [definition]. (**0.27** seconds)

ee Merriam-Webster Dictionary

eat·en \'ē-tən\; **eat**·ing. Etymology: akin to Old High German ...
Cached - Similar

his original ideas: salads made in small batches several times a day; overstuffed sandwiches on very thin, ...
static.elizabar.com/stores-locations/**EAT**/index.htm - Cached - Similar

EAT: Summary for Brinker International, Inc. Com- Yahoo! Finance
Get detailed information on Brinker International, Inc. Com (**EAT**) including quote performance, Real-Time ECN, technical chart analysis, key stats, ...
finance.yahoo.com/q?s=**Eat** - Cached - Similar

News results for **eat**

Push to **Eat** Local Food Is Hampered by Shortage - 3 hours ago
DEMAND AND NO SUPPLY Kevin McCollister raises sheep and pigs on his farm in East Montpelier, Vt., and has had trouble arranging for slaughter. ...
New York Times - 3 related articles »
Calorie Counting Applications Fuel **Eating** Disorders - ABC News
Chili's owner Brinker raises 2010 outlook - MarketWatch - 131 related articles »

Local business results for **eat** near Princeton, NJ - Change location

Time to **Eat** Diner, Inc. - www.timeto**eat**diner.com
270 Route 202/206, Bridgewater - (908) 704-9221
"The food is great and they are open late."
★★★★☆ 15 reviews, directions, hours, and more »

Eat Pray Love Movie
Enter the **Eat** Pray Love sweepstakes to let yourself go. Opens 8/13.
Facebook.com/**Eat**PrayLoveMovie
Watch trailer

Figure 7-8. Search result of “water” in the sidebar library; display media type “sound” is selected at the bottom of the sidebar.

Multiple pictures or sounds can be assigned to a concept, and are each listed in the sidebar when the concept is searched. Users can change the order of the representations (Figure 7-9 (a)) by moving a selected item up or down. This is a way to indicate a judgment on the effectiveness of the visual illustrations. Once the users save the ranking and change the display setting to “user preference,” the representations are always shown in their customized order. Users can switch back to the order provided by the system by changing the display setting back to “system default.”

Users can select their preferred picture/sound (Figure 7-9 (b)) to display in the popup dictionary instead of the default Lingraphica icon. Also, the top picture/sound in the customized ranking is automatically set as the favorite for display. The user preference is recorded in a text file for OMLA to check before loading illustrations into the popup box.

Once the logged ranking and preferences are sent to the server, they can be utilized to supervise machine learning algorithms to refine mechanisms in concept related image retrieval. Such algorithms can filter out less effective representations, and update the system order of the visual/auditory representations. Furthermore, the library can be expanded with automatically selected web images.

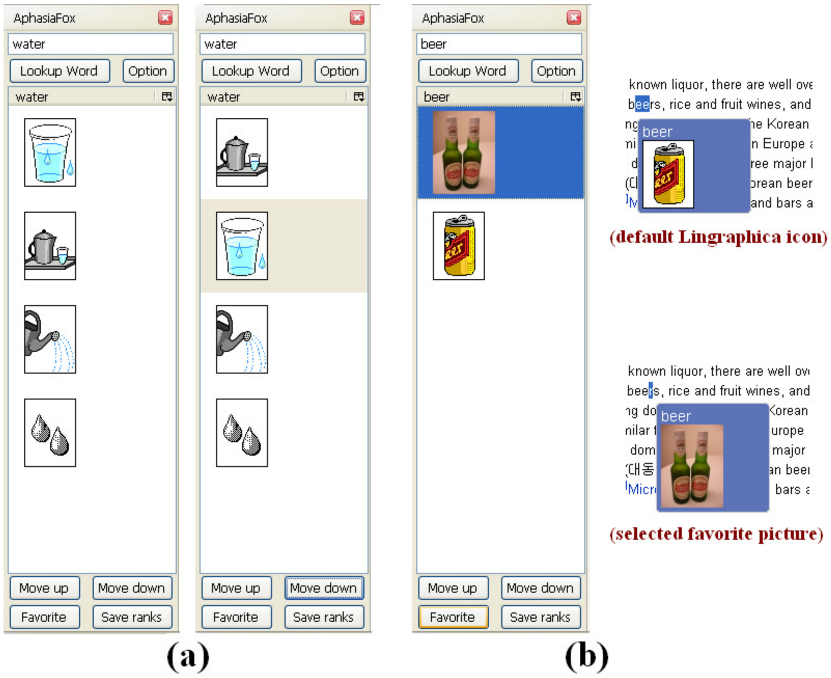


Figure 7-9. Adaptable features in the sidebar library: (a) preference ranking; (b) select favorite representation.

7.1.4.2 Multimedia Upload Dialog

Users can submit personal images for illustrating a specified concept through the upload dialog (Figure 7-10, top). Once the server receives the request, the image is compressed into a thumbnail (Figure 7-10, bottom) with a maximum height of 80 pixels. A copy of the original image is kept in a separate folder. The thumbnail is used in the popup dictionary and sidebar library display, since its small size facilitates rapid Internet data transfer. The original images are kept for potential later image processing. It is planned to extend this function to allow user-uploaded sounds and videos.

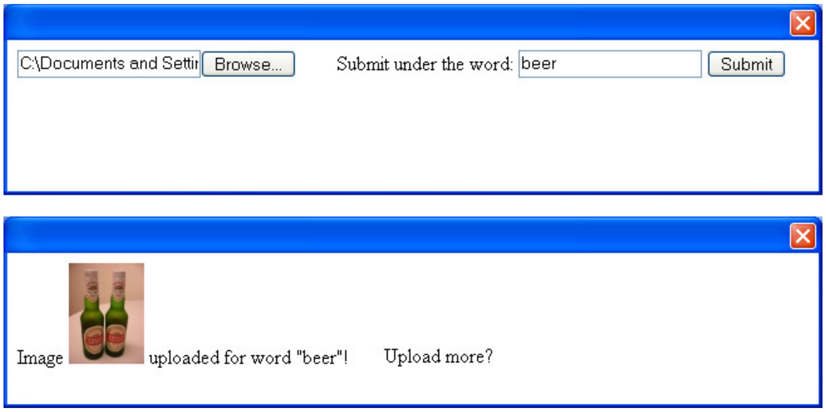

C:\Documents and Settir Browse... Submit under the word: beer Submit

Image uploaded for word "beer"! Upload more?

Figure 7-10. Picture upload dialog.

7.1.5 System Usability Evaluation

7.1.5.1 Tasks and Participants

The task list for the usability test consists of seven breakdown tasks and a scenario task: (1) Turn on AphasiaFox at a webpage for Kaiseki cuisine; (2) look up the six words in the popup dictionary and comment on the definition picture: (two dishes, two ways of preparing/cooking food, and two words in other parts of speech (prep., adj., adv., etc)); (3) upload an image for the word "sashimi;" (4) search for all the pictorial representations of the word "Kaiseki" in the library; (5) rank the pictures for "dessert" based on your own preference and save the result; (6) set a different picture as the favorite for "Hassun", and check the result; (7) change the display option so that pictures will be listed in personalized order, and check using "dessert;" (8) go to another website: The Complete List of the Beijing Olympic Sports. On the webpage, find a sport that has not been defined in OMLA, the multimedia dictionary/library. Upload three images for that sport, and set one as your favorite for display. Search for a sport with pictures already in the library that you are interested in, and re-rank the pictures if the order is different from your preference.

Nine graduate students with an average age of 24.8 and an average 10 years of web experience participated in the usability study. Six of them were non-native English speakers.

7.1.5.2 Results and Feedback

In general (based on a 1~5 point scale, 1 being the easiest, 5, the hardest), participants found AphasiaFox easy to use (2.22) and the interaction went smoothly (2.56). The key findings and recommendations from the usability study include:

(1) Six out of nine people forgot to turn on AphasiaFox when switching to a new tab.

(2) Click/select words to view the popup dictionary is easy in general. One participant preferred mouseover. Two people found the pictures too small. Two people requested to have a button in the popup box to start the sidebar from that location.

(3) Four people had trouble with the popup dictionary and the sidebar library started with different menus (one from "tool," and the other from "view").

(4) People, in general, found that setting their favorite picture was easy, though two of them confused this task with ranking. Two people forgot to save the ranking, and the Firefox "find" toolbar blocked part of the sidebar.

(5) Two people requested to have the selected word on the webpage automatically copied to the sidebar.

(6) Two people forgot to specify a word for the image to upload for the first time. Three people requested to upload multiple pictures at a time, and a wanted way to tell if the selected picture has been uploaded.

Overall, AphasiaFox, especially its popup pictorial dictionary, was found to be an easy-to-use web language facility.

7.1.6 Discussion and Conclusion

There are many things to improve in AphasiaFox. The following are some major issues that are revealed in the design, implementation, and evaluation processes:

(1) We only look for the exact match for the compound words, and thus “ice creams” cannot be recognized. The morphing of compound words is more complicated, for example “bus stopped” should not be simplified as the compound now “bus stop”; however, “bus stops” is ambiguous.

(2) Right now, we merely assume that the words are all in their most common meaning. However, a word can be a different part of speech, and can mean quite differently in various context (for instance, the word “water” in Figure 7-9 (a)). Showing a visual representation in another sense for a selected word may cause confusion, especially for language learners. Therefore, word disambiguation should be taken into consideration.

(3) At the moment, any user can upload any picture to the database. There is no quality control over the user input. Mechanisms should be implemented to monitor and filter out potential poor submission. Also, the uploaded images are assigned to a word without definition, so there is no way for the server to differentiate among various senses, although computer vision methods may help.

(4) Users can indicate their favorite picture/sound and overall preference, but are not allowed to delete or vote against a picture/sound. If users can be involved in the data quality control by marking false input from others, resulting in less burden for the server.

(5) The current OMLA system has not yet integrated the adaptive part into the application. The balance between adaptable and adaptive component could be tricky: how the user’s personalized preference can be merged with the newly updated database.

In this section, I described the architecture and design of OMLA, an Online Multimedia Language Assistant for people with language disabilities or barriers. Concepts associated with a variety of multimedia representations are organized in a customizable adaptive vocabulary hierarchy and made available over the Internet. Individuals with communication difficulty can look up words for information comprehension, communication, and language rehabilitation.

Our group is currently working on improving and evaluating the different components of the system which will be integrated into a complete functional system. We are considering enhancing OMLA with ImageNet. Now that we have set the foundations of the user-friendly online multimedia library, we will use it to build applications that will allow people with communication support needs to find and share information, and collaborate more efficiently with others.

7.2 Information Comprehension with OMLA in Food Domain

In the last section, I introduced the Online Multimedia Language Assistant (OMLA), a web-based communication support system utilizing the multimedia augmented semantic network constructed in the research discussed in the previous chapters. The popup dictionary provided by OMLA allows users to instantly look up visual or auditory representation for unfamiliar concepts that they run into when browsing webpages. A usability study showed that OMLA and its adaptable features are easy to use.

However, up to that point, whether the multimedia language can overcome language obstacles and enhance people's comprehension of information via a web platform remained unknown. To investigate how well multimedia representations can facilitate instant information understanding, two studies, one in the food domain and the other in the medical care domain, were conducted comparing web content comprehension performance of native and non-native English speakers with and without the assistance of OMLA. This section will introduce the food study, and the next section will describe the medical care study.

7.2.1 Food-related Information Comprehension, Study Motivation

Food as the testing scenario was chosen because it is representative of domains where people encounter cross-language translation issues.

In daily life, people commonly walk into a foreign food restaurant, look at the menu, and have clear no idea what to order from the menu. This is because some of the ingredients are listed in the foreign language, or only exist in that country. People might be unfamiliar with the ingredients or the way the dish is prepared. Even when others (e.g. the waiter) explain the menu further, it is still hard to imagine what the dish looks and tastes like. It is also potentially dangerous if people have food allergies or diet constraints. To help resolve the issue and to advertise, restaurants in China often have pictures printed by the name of the dishes. Similarly, in Japan, there is a custom for a restaurant to show plastic/wax models of their dishes in the restaurant window.

Similar situations happen when following cooking recipes. People may not know the materials needed and have trouble picturing how the food should turn out after each step of preparation. Terms (e.g. cooking ingredients) in one language may not have a direct translation into another, and even if they do, people may not know them or be able to picture the actual objects or actions based on the definition given.

7.2.2 Experimental Design and Hypothesis

The basic design of the study was to match dishes with different cuisine descriptions to pictures. People's performances with text-only descriptions and picture-enhanced descriptions (with assistance of OMLA pictorial popup dictionary) were compared among native and non-native English speaker groups. A stage in which only dish names were given was added as the baseline to assess participants' pre-acquired knowledge of these cuisines. Participants were asked to indicate their confidence for each match, indicating their understanding or comfort. The study is a two by two factorial within subject design. The independent variables are display stimuli (text only vs. text + picture) and participants' native language (native vs. non-native English speakers). The dependent variables are matching accuracy, confidence level, and OMLA usage (measured by the number of times people use OMLA in the text+picture mode). We are aware that there is a learning effect in the within subject design, which can be a confounding factor with the matching accuracy. A between-subject design with matched pairs would have been more appropriate. However, we hypothesized that people's food domain knowledge on different cuisines and English proficiency (related to food terminologies) could affect their understanding of the information. These two factors are hard to measure. People's self-report evaluation may not be as accurate. Also, people's knowledge about different cuisines varies greatly (even those who live together), and thus it is difficult to find good match pairs. Furthermore, keeping food domain knowledge and English proficiency as strictly controlled variables may lose some important insight into the effect of multimedia language support. Therefore, we kept the within-subject design and tried to minimize the learning effect by letting people choose to view the pictorial support. The assumption is that if a person can get sufficient clues from the text, he/she will not bother to click to see the pictures.

The principal hypotheses we had in mind included:

H1. Pictures plus text works significantly better than text only for non- native English speakers.

H2. Pictures plus text works significantly better than text only for English-native speakers, because there are terms from other languages embedded in the descriptions of the dishes.

H3. Pictures plus text works significantly better for non- native English speakers than for native speakers.

H4. People are more confident with their understanding of the information given pictures.

H5. Non-native speakers need more assistant from OMLA than native speakers.

H6. Knowledge about food in different cuisines is a significant influential factor.

7.2.3 Data Preparation

7.2.3.1 Selection of List of Dishes

The criteria for dish selection were: 1) popular/representative within the cuisine; 2) distinctive, ingredients or features (e.g. color and shape) can differentiate it; 3) cannot be fully identified by dish name; 4) ingredients are unique to that cuisine or have a foreign name. Thus, some of the dishes can be easily mistaken for another dish, and people need to know the dish well or fully understand the description to match the picture (e.g. Manhattan vs. New England clam chowder).

The 21 dishes picked, with the first “chocolate truffle” as a demo, covered a wide range of food types (Table 7-2). They are among the most popular dishes from cuisines in three categories: 1) American food or food with a foreign origin, but well-known to most Americans, such as “pizza;” 2) Foreign cuisines common in the USA, such as Italian, Mexican, Indian, and Thai; 3) Cuisines common in the US, but the dishes selected are common in the home countries but not yet in the USA, such as French and Japanese; 4) less common foreign cuisines (Singapore). Dishes 1-3 are popular in the USA, or at least people can figure out what type of food they are even if cannot disambiguate within its category. In this way, participants would not be intimidated by encountering dishes completely unheard of. Dishes 4-20 were organized in alphabetic order.

	Dish Name	Cuisine	Category
0	Chocolate Truffle	French?	Dessert
1	Chicago Style Pizza	American	Batter
2	Tamale	Mexican	Appetizer
3	Manhattan clam chowder	American	Soup
4	Brandade de Morue	French	Appetizer
5	Crème Caramel	French	Dessert
6	Cannelloni	Italian	Pasta
7	Daal Makhani	Indian	Soup
8	Goi Cuon	Vietnamese	Appetizer
9	Ice Kachang	Singapore	Dessert
10	Katsudon	Japanese	Rice
11	Mixed Satay	Indonesian	Meat
12	Nasi Lemak	Indonesian	Rice
13	Oden	Japanese	Hotpot
14	Okonomiyaki	Japanese	Batter
15	Paratha	Indian	Bread
16	Phad Thai	Thai	Noodle
17	Pot au Feu	French	Meat
18	Pumpernickel Bread	German	Bread
19	Ratatouille	French	Vegetable
20	Somen	Japanese	Noodle

Table 7-2. Dishes, cuisines, and categories

Descriptions of the dishes came from online and paperback recipe books, with several versions compared and integrated. Descriptions contained the most distinctive features and ingredients of the dishes. Foreign names of the ingredients were deliberately kept for testing purposes.

7.2.3.2 Selection of Picture Representations

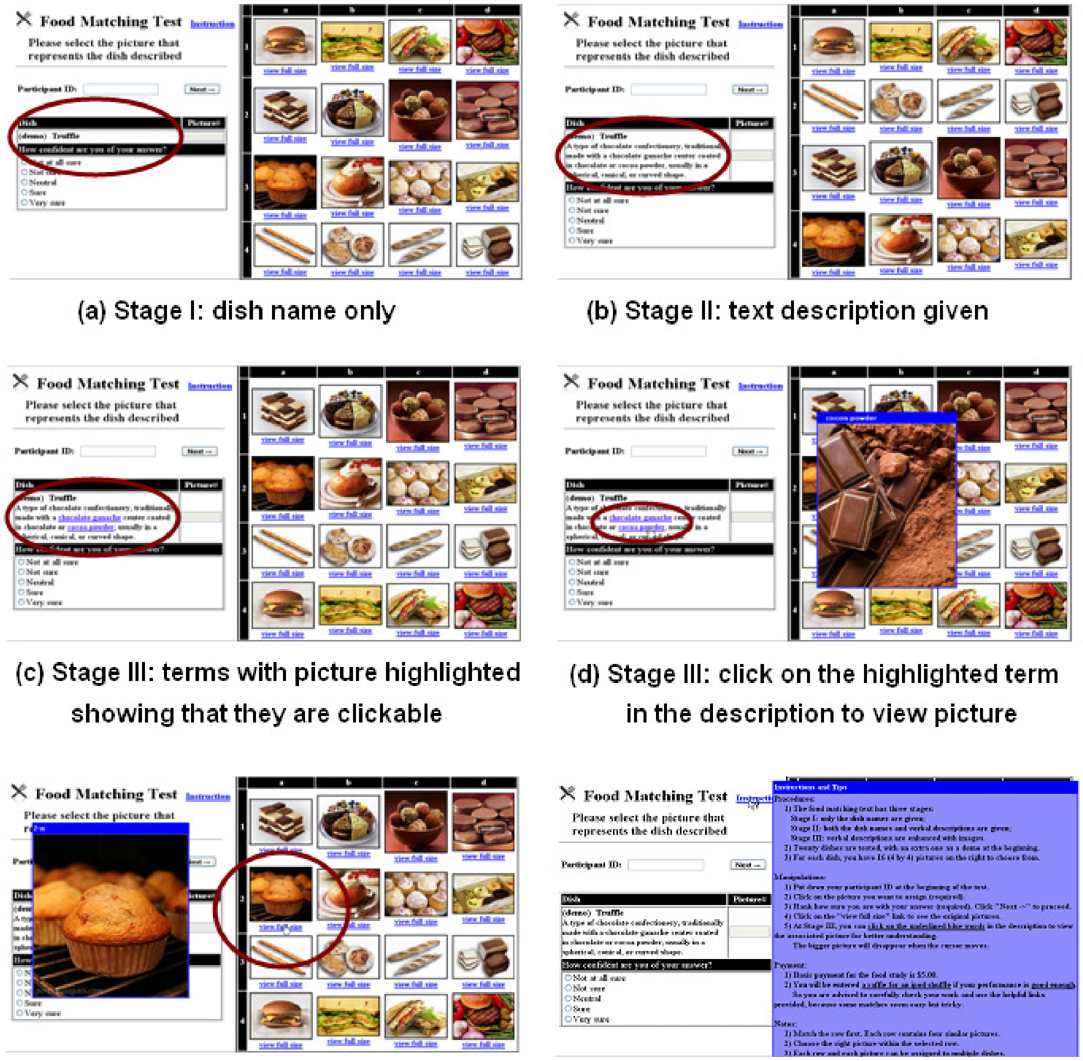
Pictures of the dishes and ingredients were collected from online recipes, blogs and posts about cooking, albums from cooking lessons, and image search results. Although the picture sizes were not unified, no pictures with low resolution were included. For dish pictures used as the options, both the thumbnails created by OMLA during uploading and the original-sized ones were kept and used. Three people checked the pictures to ensure that the pictures introduced into OMLA are good illustrations of the actual dishes/ingredients, as well as that the other pictures used as options were not too similar to the intended picture to distinguish by the given information. Dish descriptions were further tailored according to the pictures as well. Ingredients that did not show up in the pictures were removed so as not to be confusing during the ingredient look-up.

7.2.4 Interface and Procedure

7.2.4.1 Study Interface

The study interface is web-based. All the materials and interactions are located on the same page, so no scrolling is involved. On the left are the names (and descriptions) of the dishes, the assigned picture ID, and the ratings indicating the participant's confidence in the picture selection. On the right is a matrix of 16 thumbnails of pictures to choose from. Similar items were put in the same row (e.g., a row contains all chocolate items and another row contains all bread items). This setting reduced the workload of participants so that they did not need to memorize positions of pictures for comparison. Figures 7-11 (a)(b)(c) show the snapshots of the study interface at different stages (highlighted in red circles): Stage I – matching with dish names only; Stage II – text descriptions added to help with the matching; and Stage III – certain terms in descriptions enhanced with pictures as illustrations of the concepts.

The Online Multimedia Language Assistant was applied in Stage III with modifications. Words and phrases detected in the OMLA vocabulary were colored and underlined so that people knew what to click on to see an associated picture (Figure 7-11 (d)). This avoided the frustration of clicking on many words and receiving a "not in dictionary" alert. People could click on the "view full size" link under each option picture to see the original picture (Figure 11 (e)). This way, details of the dishes could be better reviewed and compared. There was another "Instructions" link on the top by the title that people could click to refer to the instructions and explanations on study stages, manipulations, payments and notes required to go over at the beginning of the study.



Please select the picture that represents the dish described

Participant ID:

(demo) Truffle

A type of chocolate confectionery, traditionally made with a chocolate ganache centre coated in chocolate or cocoa powder, usually in a spherical, conical, or curved shape.

How confident are you of your answer?

○ Not at all sure
○ Not sure
○ Neutral
○ Sure
○ Very sure

Instructions and Tips

Procedures:
1) The food matching test has three stages:
Stage I: only the dish names are given;
Stage II: both the dish names and verbal descriptions are given;
Stage III: verbal descriptions are enhanced with images.
2) Twenty dishes are tested, with an extra one as a demo at the beginning.
3) For each dish, you have 16 (4 by 4) pictures on the right to choose from.

Manipulations:
1) Put down your participant ID at the beginning of the test.
2) Click on the picture you want to assign (required).
3) Rank how sure you are with your answer (required). Click "Next ->" to proceed.
4) Click on the "view full size" link to see the original pictures.
5) At Stage III, you can click on the underlined blue words in the description to view the associated picture for better understanding.
The bigger picture will disappear when the cursor moves.

Payment:
1) Basic payment for the food study is $5.00.
2) You will be entered a raffle for an ipod shuffle if your performance is good enough.
So you are advised to carefully check your work and use the helpful links provided, because some matches seem easy but tricky.

Notes:
1) Match the row first. Each row contains four similar pictures.
2) Choose the right picture within the selected row.
3) Each row and each picture can be assigned to multiple dishes.

(a) Stage I: dish name only

(b) Stage II: text description given

(c) Stage III: terms with picture highlighted showing that they are clickable

(d) Stage III: click on the highlighted term in the description to view picture

(e) view enlarged option picture

(f) review instructions

Figure 7-11. Interface for the food matching study, in three stages.

7.2.4.2 Study Procedure

The procedure of the study was as follows: after going over the dish name and descriptions (in Stage II and III), participants clicked on the picture on the right that they want to assign to the dish (no typing). Participants could first narrow down their selections to a row (broader type of food) and then choose from within the row. After making a choice, participants needed to select their confidence level (from “not at all sure” to “very sure”). Clicking the “Next →” button submitted their answer and proceeded. Confidence ratings let participants indicate whether they had no clue about the dish and were just guessing or if they actually felt they knew the correct match.

The set of 16 pictures to choose from for each dish was predetermined and different from each other. The option pictures may be from the same cuisine or same category as the intended dish picture. The position of the rows was randomly assigned each time the webpage was reloaded, so at each stage the same pictures appeared for a dish but at different positions. This made participants go over choices again instead of just remembering the position of the selected picture from the last stage, to minimize the possibility of skipping correct pictures over and over again.

The interface automatically logged the matched picture, confidence rating, which option pictures were viewed in full size, as well as which words in the descriptions were clicked to see the associated picture (Stage III), indicating if people actually used the pictures in Stage III matching.

7.2.5 Participants

Ten native and ten non-native English speakers participated in the study. Within the native speakers' group, half of them were female. One person was younger than age 20, three people between 21 and 30, three people between 31 and 40, 1 person between 41 and 50, and two between 51 and 65. All of the participants had an education level of some college or above. Within the non-native speakers' group, there were four females. Two people were between the age of 31 and 40, and the rest fell into the 21 to 30 span.

The native languages of the non-native group are listed in Table 7-3. Since most of the unknown foreign ingredients were from Japanese and South East Asian cuisines, subjects were not recruited from these regions. Most native speakers knew multiple languages at different levels (e.g. Spanish, French, etc.). This was less common among non-native speakers. The number of foreign languages and speaking ability did not significantly impact the study's results.

Native Language	Number of Participants
Mandarin (Chinese)	6
Cantonese (Chinese)	1
Russian	1
German and French	1
Tamil (Indian)	1

Table 7-3. Non-native English participants' native languages

All participants' English level were measured with a rating scale (1: very poor; 2: poor; 3: fair; 4: good; 5: excellent) based on their self-assessment. All native speakers obtained a 5, two non-native speakers got 5, six people got 4, and two got 3. This is only an estimate of people's general English skills, not specific to their comprehension of food terms.

People's knowledge of food domains was evaluated (0: no idea at all; 1: just a little; 2: some; 3: a fair amount; 4: quite a lot; 5: expert) as well, according to their experience with cooking and dining in different cuisines as well as their ways of learning to cook and picking dishes from a menu. This provided some hint as to their attention to different dish names and ingredients. However, it is only a rough estimate. Observation in the study showed that sometimes people are more/less knowledgeable in certain cuisines than they have indicated.

7.2.6 Results and Analysis

Results from the food matching experiment verified our hypothesis that both native and non-native English speakers benefited significantly in information understanding from pictures in food domains. The improvement received from the assistance of pictures was greater for non-native English speakers than native speakers. Both level of English and knowledge about food and different cuisines were essential influential factors. Test of homogeneity of variances showed that native and non-native English speakers' matching performance across different stages came from the same normal distribution. It indicated that these conditions were matched. Therefore, an analysis of variance was used to analyze the data.

7.2.6.1 Text Only vs. Text + Pictures, Comparison across Stages

Figure 7-12 (left) showed the average improvement in terms of number of correct matches and confident correct matches (participants rated as "sure" or "very sure") over the three stages of the experiment (Stage I: given only dish names; Stage II: text only descriptions provided; Stage III: text descriptions enhanced with pictures). The latter eliminated correct matches achieved by pure guessing. The improvement is significant (for correct matches: $F(2,54) = 67.269$, $p < 0.001$, $\eta^2 = 0.589$; for confident correct matches: $F(2,54) = 79.359$, $p < 0.001$, $\eta^2 = 0.673$). The progressively diminishing difference between correct matches and correct matches with confidence indicates that people applied fewer guesses and understood more about the dishes.

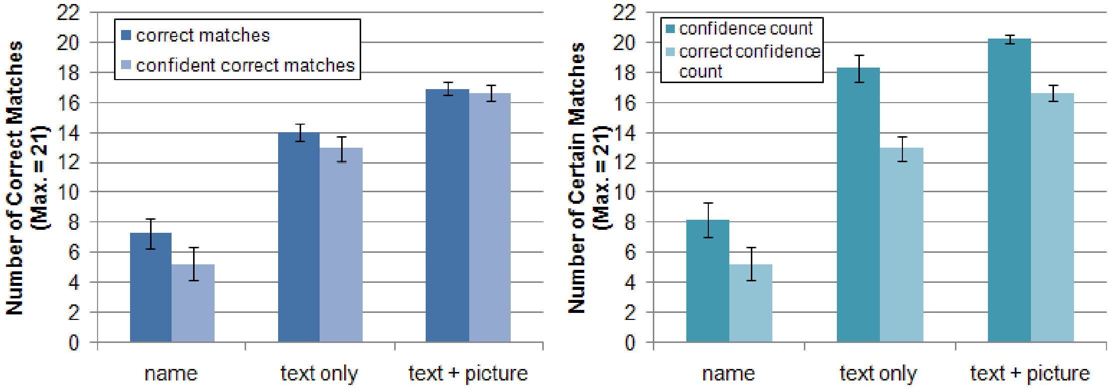


Figure 7-12. Comparison of number of correct matches (left) and number of confident match (right) across different stages.

Figure 7-12 (right) shows the change of confidence level in terms of how many matches and correct matches were rated as “sure” or “very sure” across the three stages. The differences are significant, F(2,54) = 72.793, p < 0.001, η^2 = 0.704 for the former and F(2,54) = 79.359, p < 0.001, η^2 = 0.673 for the latter. The large confidence increase from Stage I to Stage II showed that information was obtained through text descriptions. The confidence increase in Stage III may be from the aid of pictures or comprehension reinforcement by just rereading the descriptions.

7.2.6.2 Native vs. Non-Native English Speakers, Comparison across Participants

A further look at the difference in performance between native and non-native English speakers at each stage is shown in Figure 7-13 (left). The difference is significant between the two groups in number of correct matches with guesses eliminated (F(1,54) = 20.743, p < 0.001, η^2 = 0.088). Overall, native speakers got significantly more matches right at each stage than non-native speakers, but the difference decreased with more information provided. However, the interaction between participant groups and stages was not significant (F(2,54)=1.181, p=0.315, η^2 =0.010), implying similar improvements for both native and non-native speakers.

In Figure 7-13 (left), the columns on the right noted as “picture” show the number of confident successful matches for which the pictures in the descriptions were clicked and viewed in Stage III. They were plotted to be differentiated from those determined completely based on previous knowledge and things learned from the verbal descriptions in Stage II. Although at each stage native speakers did better, significantly more correct matches were achieved with the help of pictures for non-native English speakers than for native speakers (F(1,54) = 5.942, p = 0.018, η^2 = 0.043). Figure 7-13 (right) shows the number times people clicked on the words to view associated pictures in OMLA. On average, non-native speakers clicked on pictures 50 times more, indicating that they used more assistance from the pictures than native speakers (F(1,18) = 10.248, p = 0.005, η^2 = 0.363).

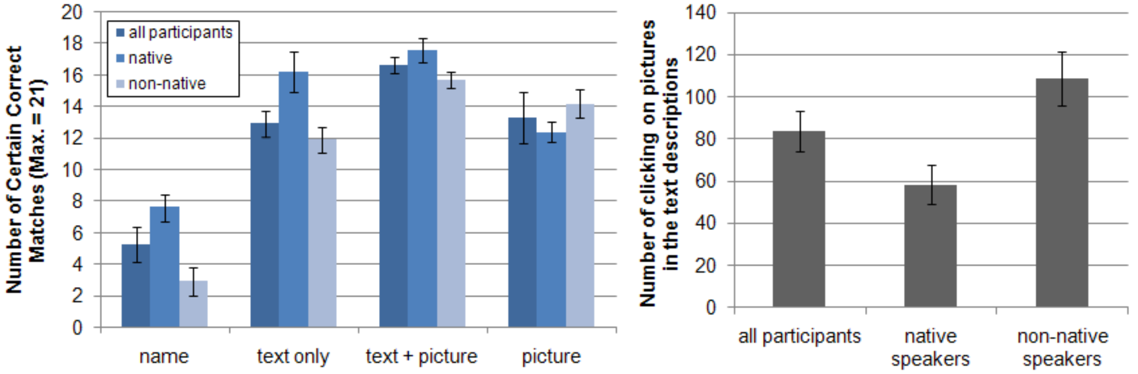


Figure 7-13. Comparison of number of confident matches (left) and number of clicking to view OMLA pictures (right) between native and non-native English speakers.

7.2.6.3 Post-hoc Test, Influence of English Proficiency

Even within non-native speakers, their proficiency levels of English, especially vocabulary, varied a lot from person to person. Actually, the performance of two non-native English speakers whose English proficiency was marked as "excellent" was quite similar to that of the native speakers. It suggests that English proficiency, whether English is the native language or not may be an influential factor.

Figure 7-14 (left) shows the performance differences affected by English proficiency for both native and non-native speakers at Stage II and III, as well as the number of correct matches with the help of pictures, which is significant (F(1,18) = 15.636, p < 0.001, η^2 = 0.465). At Stage II with only text description, people with better English and bigger vocabularies did better in the matching tasks (F(1,18) = 31.976, p < 0.001, η^2 = 0.640). At Stage III, people with fair English browsed pictures for almost every dish they got right, while people with excellent English only clicked on pictures for about 2/3 of the matches they got right. The total number of times people checked OMLA was significantly different (Figure 7-14(right), F(1,18) = 14.065, p = 0.001, η^2 = 0.439). On average people whose described their English as "fair" viewed pictures nearly 100 times more for all the 21 dishes than those considering their English as "excellent."

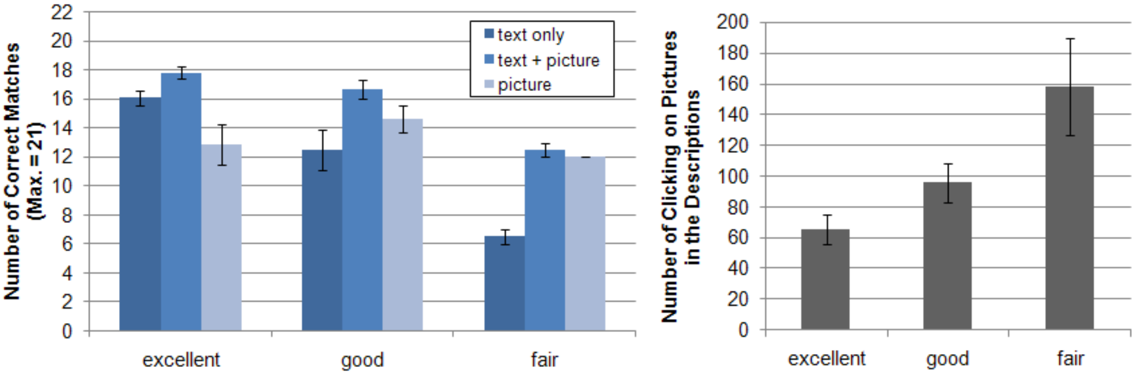


Figure 7-14. Comparison of number of confident correct matches in Stage II and III, as well as confident correct answers using OMLA in Stage III (left); number of clicking to view OMLA pictures in the descriptions at different English proficiency level (right).

7.2.6.4 Post-hoc Test, Influence of Food Domain Knowledge

Another factor that can greatly impact people's food matching performance is their knowledge about food, more specifically different cuisine and ingredients. For example, many of the Chinese participants only or mainly cook Chinese food and dine at Chinese restaurants.

As shown in Figure 7-15 (left), the number of correct matches decreased significantly as the amount of food knowledge went down, especially in Stage I (F(1,18) = 188.560, p < 0.001, η^2 =

0.913) and Stage II (F(1,18) = 37.336, p < 0.001, η^2 = 0.673). On average, at the name-only stage, people who knew the tested cuisines quite well got about 14 matches right, while people with little knowledge were successful in only about three dishes. At the last stage, even with text and pictures, those who were unfamiliar with different cuisines made about six mistakes. However, their improvement compared to Stage II was larger than knowledgeable subjects. They also counted on the pictures significantly more (Figure 7-15 (right), F(1,18) = 8.975, p = 0.008, η^2 = 0.333).

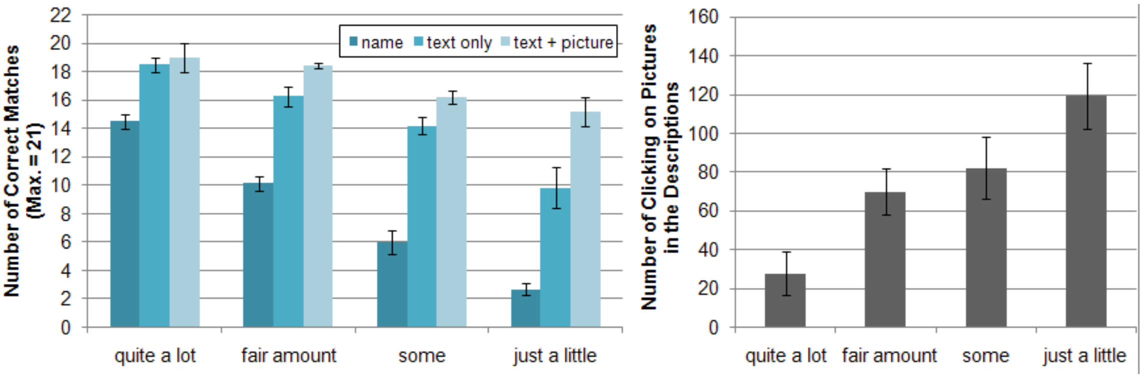


Figure 7-15. Comparison of number of correct matches among participants with different level of food knowledge (left); number of clicking to view OMLA pictures in the descriptions with different amount of food knowledge (right).

7.2.6.5 Subjective Feedback

Participants' comments on the study were collected. Eighteen of the twenty participants described the study as a lot of fun. Half of the participants rated the study difficulty level greater than or equal to 4 (1=extremely easy, 5=extremely difficult). Four people (all native speakers) thought it was harder than they initially imagined. The difficulty came from 1) unknown foreign dish names and ingredients (20 people mentioned this); 2) lack of familiarity with the cuisines (seven people); 3) the need to check details and compare across multiple pictures (four people).

All of the participants agreed that in general the pictures were very clear and helpful, especially for unknown terms. Two people were "amazed by how much easier it was with text and pictures together." The exception (mentioned by several) was pictures for various cooking sources (e.g. soy sauce), which cannot be well explained from pictures of a completed dish.

Overall, people exhibited no difficulty in using the OMLA popup pictorial dictionary. One person mentioned a desire to have two full-sized pictures showing at the same time for comparison, and another person wanted to have the text and pictures sitting side by side.

7.2.7 Discussion and Conclusions

There were some interesting findings in our study, and participants' post-study feedback:

(1) Unfamiliarity with the cuisines and dishes came partially from diet. Two participants were vegetarian/vegan, and commented that many of the dishes they had never and will never eat. They also mentioned that they would need to check with the waiter carefully if they were not sure about the ingredients of the dishes when going to a restaurant. This was also true for people with food allergies, especially when they were ordering from foreign cuisines.

(2) People with relatively poorer English failed in the matching tasks because of other English terms in the descriptions that had no associated OMLA picture. For example, one subject checked hotpot-type dish pictures over and over again while the given dish was stir-fried noodles. They also experienced information overload especially when the description was long and had many unknown ingredients.

(3) Some generic terms are hard to illustrate with a single picture. For example, the word "stew" was assigned to a picture of a pot of beef stew, and people got confused when they checked the word "stew" in a vegetable stew dish and saw meat in it. One participant suggested a grid of four different stew dishes (showing the motion of stirring) at the same time for "stew."

(4) When some people were sure of one match (even if it was wrong), they did not bother going over it again in the next stage. This situation was extremely severe in the pilot study. Giving feedback on errors for each stage in the actual study helped a lot with encouraging people to be more careful and check their work, but there were occasional lapses.

(5) When people felt that they knew a word well, they would not click to see the picture, even if actually viewing the picture could assist the comparison, especially when the word denotes the key characteristics of the dish, such as "deep dish" in "Chicago-style pizza." It was often seen that people checked pictures for all "strange words" many times but still had no clue about the dish. Some of them finally clicked on "normal words" to receive the hints.

The points listed above can help us better understand how people perceive information, which can further guide us in modification of the Online Multimedia Language Assistant in both pictorial representation selection and the interaction setting, so that people benefit more from the system.

To sum up, a study in the food domain, which is a typical scenario with many unfamiliar foreign terminologies, was conducted to evaluate how well pictures can enhance verbal information comprehension compared to pure text with both native and non-native English speakers. Results confirmed our hypothesis that pictures are significantly helpful for understanding information across language barriers. Both the level of English and domain specific knowledge (food in this case) had significant effects on the results.

7.3 Information Comprehension with OMLA in Medical Care Domain

Section 7.2 presented a study that evaluates the effectiveness of the Online Multimedia Language Assistant on enhancing food domain information comprehension with both native and non-native English speakers. Results showed that multimedia representations can compensate for the insufficient language skills and promote people’s understanding of information provided by the Internet. To further validate the generalization of this finding, another study was in a different domain, i.e. medical care. Although the purpose is the same, the two studies applied different methodologies given the differences in the two tested domains. Given the difficulty of depicting medical terminologies with pictures, a medical picture labeling study was carried out on Amazon Mechanical Turk prior to the main experiment. This section starts with justification of the choice of scenario, and then move on to the detailed study description.

7.3.1 Communication Barriers in Medical Care and Conventional Approaches

Even in countries with modern medical facilities, sufficient doctors and easy access to medical care, problems exist in providing this care because patients fail to provide important information about their illness to medical personnel, and fail to follow instructions for care (Gazmararian et al.1999). There can be many reasons for these failures. They include language and cultural barriers, low literacy rates, poor knowledge about basic medical practices and common diseases, and a lack of knowledge about regional medical issues. In addition, patients are often anxious when visiting a doctor because of their illness, so that important questions and instructions are missed in the verbal exchange.

Illness process and medical encounter is a comprehensive sequence starting with self-awareness of uncomfortable body feelings and changes that are eventually labeled as “sick” or “ill.” Steps, from self-treatment, family care, to professional consultation are then taken to remove symptoms or recover from the illness. In this process people may encounter “medical literacy” problems with the involvement of a legitimate health care system (Kleinman et al.2006). For example, in a university, when international students or a member of their family become ill, they may go to the student health center for advice. Communication is compromised in these situations because the international patient may have different perceptions of their illness, suffer embarrassment, and/or lack knowledge of the proper steps to take to ameliorate symptoms. This raises complications in the patient describing his/her condition and understanding directions for the prescribed medications, the diagnosis and overall treatment.

Confidentiality concerns diminish or eliminate the use of third-party interpreters in this exchange, and it is often inconvenient to bring a friend or family member to a clinic. Medicine is also conducted differently in other countries. Senior citizens who travel or immigrate to a new country are used to treatments in their home country and have difficulty accepting and adapting to the methods used in the new country's medical culture. Younger people who visit other countries for education, work, business, and tourism are not used to getting medical care and are often unaware of the medical terms and medical solutions used in their own country. Just building trust in those situations where medicines and/or treatments are given by a person of a significantly different culture is a huge threshold to overcome.

Ordinary translation dictionaries are not useful in these illness treatment circumstances because most of them do not translate complex medical terms, and, even if they did, the ill person may not know the term for either the symptom or the illness in his or her own language (Ley1982). Furthermore, due to cultural differences, people may not always divide body parts in the same way, and thus could lead to misinterpretation.

7.3.2 Pictorial Support in Medical Care Communication

Pictures have been proven to have positive effect on promoting communication in health care (Houts et al.2006)(Katz et al.2006). Pictures have been used to: 1) draw attention to medical information. For example, it has been found that people are more likely to read and follow handouts with a cartoon embedded in the text (Delp and Jones1996); 2) improve understanding of health information. For example, it has been shown that pictures (e.g. (Austin et al.1995)(Sojoumor and Wogalter1998)(Morrow et al.1998)), animations and videos (Leiner et al.2004) help people to understand complex medical terminology; 3) provide cues for recalling information. For example, it has been found that people have difficulty remembering a doctor's instructions (Ley1982) and that pictures help (Delp and Jones1996) (Morrow et al.1998).

The studies listed above also raised other important issues about the use of pictures. For instance, the style of pictures (Delp and Jones1996) (photos vs. colored drawings vs. black and white line drawings), and the content of the pictures (Morrow et al.1998) (pictures of an entire person vs. pictures of a body part) may affect their comprehension.

The approach proposed here differs from the above prior work in five ways:

(1) A computer system was used to present the pictorial representations. Prior studies used paper-based handouts and were not interactive. This limits the presentations to pre-defined roles and also limits the media to be used, e.g., animations and sounds were not available.

(2) Prior studies focused on educational material and care instructions. This study focused on the entire patient care system, both diagnosis and treatment.

(3) Prior studies did not determine whether the “best” images were used. Here, multiple images and animations were evaluated prior to the study in order to select the ones whose meaning is best understood.

(4) Prior studies used a low medically literate population. We focused on building a system that combines verbal and pictorial support allowing those who have high literacy to obtain their medical information quickly through text alone.

(5) The Amazon Mechanical Turk (AMT2009) was used as a rapid method for collecting efficacy measures on our images as opposed to time consuming interviews.

7.3.3 Study Motivation and Overview

Pictures have been shown to be highly successful in communicating medical instructions to patients with low medical literacy in a host of studies. They have also been shown to be highly effective in convincing people to adopt healthier behavior. However, giving instructions on what a patient needs to do after leaving the medical care facility does not solve the entire communication problem that starts with the entrance of an individual into medical care. I therefore took a different angle, i.e. investigating if pictures can enhance people’s understanding of diagnosis of ailments and medical treatments. It is easier to create pictures to describe to a patient that he or she must take 3 pills a day, one before each meal, with water, than to create pictures that investigate the subtle nuances given in the descriptions of a symptom.

A significant disadvantage of pictures is their lack of precision. Thus, one person may perceive a picture of someone represented to have a fever as someone who is tired or having heatstroke. It is clear that diagnosis pictures need to be tested for their precision in conveying the correct meaning even across language and cultural barriers. This is a huge task facing the designers of pictorial medical diagnosis support systems. To facilitate us in selecting pictures and putting together combination of pictures, two studies were carried out to determine what types of pictures to use and also whether symptoms and integrated medical conditions can be effectively conveyed through pictures.

The first study involved the use of Amazon Mechanical Turk (AMT). AMT allows experimenters to put up tasks for anyone on the web to do for payment. Participants were paid to tag a set of drawings or animations. This allowed us to see which drawings/animations were identified correctly and to determine what properties of the drawing or animation led to this accuracy.

The second study was more complex. It was done to generate a proof of concept, that is, to show that the drawings/animations were better than simple text descriptions of multiple medical symptoms and that the OMLA system is beneficial to people with limited language skills. Much work has been performed on patients' understanding of medical instructions. Most of these studies involved interviews which investigated what was recalled or understood about the text or pictorial instructions that were given to the patients. Pictorial representations were found to work better for people with low medical literacy (Houts et al.2006)(Katz et al.2006). For the purposes of the research of this dessertation, I needed a faster quantitative measure for comparing multiple pictorial representations of symptom combinations than that used in these prior studies.

Doctors perform diagnoses by asking multiple questions. Some of them may not seem to be related to the current problem. For example, doctors may ask about the family history of a particular disease, recent travel history and even the progression of the symptoms. The patient may not understand the relationship of these questions to the problem and may omit important diagnostic information. For example, Lyme disease starts with a characteristic ring rash, but patients may only be seeing the doctor because they feel tired, are running a low fever and suffer from joint aches and pains. They may not understand what the doctor means by a ring rash and what it has to do with their physical feelings, unless the doctor shows them a picture of what is meant and how the symptoms are possibly interconnected.

Therefore, it is important to test a patient's integration of multiple pictorial symptoms into a potential diagnosis, obviously not to make patients into doctors, but to see if this helped the patient better understand the symptoms' inter-relationships and their association with possible medical treatments.

To do this, a matching study was designed. In the study participants were presented with a medical treatment for which they were to pick the illness or medical problem the treatment alleviated or cured. The illness or medical problem could be further expanded by the participant into a list of typical symptoms, e.g., a flu came with fever, general aches and pains, a sore throat and possible stomach upset. Similarly, the treatment could be decomposed into those symptoms the treatment targeted. The success of the matching therefore reflects whether people understand the medical problem in its entirety not simply the individual pictures. Of course, the matching study is not as good as a real diagnosis taking place on real patients.

7.3.4 Precision Evaluation on Pictorial Representations for Medical Terminologies

Because medical diagnosis is a complicated process, and because building such a system requires thousands of pictures, videos and animations, quick ways need to be developed to accumulate a large number of medical images and to test the images before incorporating them into the Online Multimedia Language Assistant.

7.3.4.1 Generation of Lists of Symptoms, Ailments, and Treatments

Ailments and symptoms targeted in the system came from a list provided by the Princeton University Medical Center. Table 7-4 lists (not in any particular order) the chief complaints of individuals using the medical center. This list came from 30,000 individual visit records. Eleven illnesses were picked from the list (Table 7-4 in bold format), and another 14 ailments were added that are also common but occur less frequently in younger populations, e.g., “diabetes” and “heart disease.” We deliberately chose several illnesses that have similar symptoms to see if people can distinguish them based on the text descriptions and pictorial representations.

Abdominal pain	**Gastroenteritis**	Morning after pill
Acid reflux	Pap test	Urinary tract infection
Immunizations	Birth control / contraception diseases	
Conjunctivitis	Screening for sexually transmitted	
Alcohol abuse	Mononucleosis (“mono”)	
Warts and **acne**	Other rashes like **poison ivy**	
Low back pain	Headache	**Ankle sprains**
Influenza	URI (upper respiratory infections) “**colds**”	
Anxiety	**Depression**	Fatigue
Asthma	Bronchitis	Allergic rhinitis “**hayfever**”
Sinusitis	Pharyngitis / tonsillitis (“sore throat”)	
Cough	Viral syndrome (fever without localized symptoms)	

Table 7-4. Twenty-nine most common diagnoses numerically from a university health center.

The list of medications and treatments came from two sources. First, two volunteers listed the common over-the-counter drugs that they had at home. Second, the main treatments were looked up for the ailments and symptoms collected. Table 7-5 shows the 18 treatments used in the study.

Clearasil	Amoxil	Calamine Lotion		Dramamine
EpiPen	Ex-lax	Gatorade		Insulin
Open Reduction and Internal Fixation				Pepto bismol
ProAir HFA	RICE (PRICE) protocol			Prozac
Relenza	Statin	Tecnu	TUMS	Zyrtec

Table 7-5. Eighteen treatments used in the matching study.

7.3.4.2 Selection of Pictorial Representations

There are many medical picture resources online. MedlinePlus by National Institutes of Health (NIH) (MedlinePlus2009) and drugs.com (drugs2009) both provided a medical dictionary which contains illustrated articles on all sorts of health-related topics. Their medical illustrations primarily come from A.D.A.M., an online health care information and technology company. The A.D.A.M. Medical Encyclopedia (ADAM2009) includes over 4,000 articles with an extensive library of more than 2,000 original illustrations in eight categories: disease, injury, nutrition, poison, surgery, symptom, test, and special topics. The professional illustrations focus on detailed pathological or anatomical portrayals. Without certain knowledge, it is hard for people to associate them with daily phenomena. The Hardin Meta Directory of Internet Health Sources (hardin2009) from the University of Iowa has a picture gallery of web medical images. It is an example of online medical image repositories. MedicineNet.com (medicinenet2009) obtains health and medical information offered by doctors. Together with the articles are slideshows of pictures that show different aspects of the topics, such as images of viruses and photos of symptoms. Those pictures are mainly realistic photo representations of pathological conditions that could be disturbing to look at. The icons our system used are like clip art, with a comic style representation this is hopefully easier for people to accept.

Pictures that that depicted medical terminology were collected from the Internet. The quality/size of the illustrations varied from simple comic book-like black and white line drawings, to colored clip art, to detailed illustrations from MedlinePlus (Figure 7-16). Overall, 340 drawings illustrating 250 medical terms and phrases associated with the 25 illnesses were gathered.

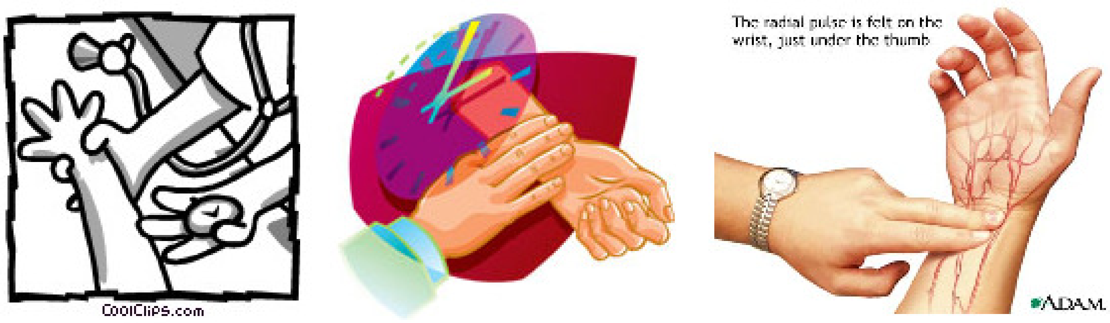


Figure 7-16. Three kinds of icons (left to right: black and white simple line drawing/cartoon, colored line drawings/clipart, and realistic illustration) for “pulse.”

Slideshows (also called animations) and videos have also been widely used in medical information demonstration, and many such directories can be accessed over the Internet. The University of Pennsylvania Health System (derived from A.D.A.M.) (penn2009) and University of Texas Medical Branch (texas2009) have slideshow collections demonstrating the reactions of organs and human body structures to diseases, and medical procedures used for diagnosis and treatment. Those slide shows are systematic depictions with speech explanations and text descriptions. Medindia Animations / Interactives (medindia2009) offers an interactive slideshow with text narration which can be played automatically, but only in the same scope. These require specific software and several minutes to play. HealthCentral 3D medical animation library (healthcentral2009), and Nucleus Medical Art (nucleus2009) are all providers of 3D animations of various health conditions. These reproduce organism functions and pathologies from a more microcosmic perspective.

Unlike the animated slideshows mentioned above, OMLA applies simple gif cartoons, which are snapshots of causes, symptoms, and treatments of various medical conditions. All animations portrayed the major symptoms of each illness. Other aspects shown in the animations which varied from illness to illness included:

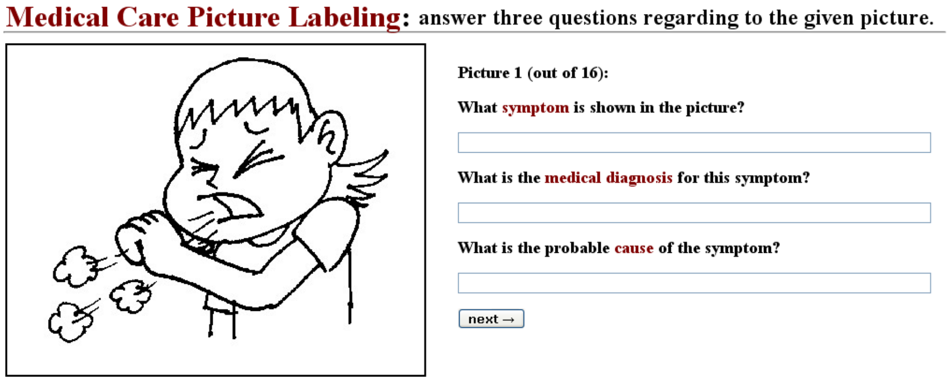
(1) External causes of the illnesses such as pollen for hay fever, (20 out of 25 animations).
(2) Infection vector, showing how people get infected, such as the spread of cold germs through contact (10 out of 25).
(3) Internal causes of a disease and its progress within the human body, such as the blockage of the coronary arteries for cardiac disease (8 out of 25).
(4) Suggestive treatments for an illness, such as quick-relief inhalers for asthma (3 out of 25).

Each animation is 5 seconds in length and contains about 10 frames. These play automatically when opened and loop once.

7.3.4.3 Verification of Representation Precision

An online study was conducted using Amazon Mechanic Turk to obtain human-generated labels for the pictorial representations collected from the web. The goals were 1) to verify if people can identify the medical concept based on the icons and animations; 2) to help us pick the best representation for the chosen medical terms and phrases; 3) to suggest rules for selection and creation of medical icons and animations. Less relevant representations were filtered out, yielding 185 drawings (40 black and white line drawings, 107 clipart files, and 38 illustrations) and 25 animations used in the study. The medical picture tagging study had 156 participants. Each representation was labelled by 10 to 20 people.

Figure 7-17 shows the web interface for the labelling study. For each run, 10 icons or animations were displayed. Participants were asked to describe the symptom(s) depicted in the picture (the tagging portion of the task), identify the medical diagnosis, and explain the possible cause(s) of the symptom(s)/ailment. Their answers (usually in short phrases or sentences) were recorded by the interface. People's justification of medical diagnosis and causes helped to understand the reasons for bad tags.



Medical Care Picture Labeling: answer three questions regarding to the given picture.

Picture 1 (out of 16):

What symptom is shown in the picture?

What is the medical diagnosis for this symptom?

What is the probable cause of the symptom?

Figure 7-17. Interface of the medical icons/animations labeling study.

No significant difference was found in accuracy rates for different representation styles. About 87% of the representations had an accuracy rate greater than 50% (Table 7-6). In general, people were good at recognizing pictures for pain, panic, fatigue, heart disease (high blood pressure, heart attack), and common flu/cold like symptoms (runny nose, fever, headache, and sore throat).

Accuracy	Count
1.0	33
0.9~1.0	17
0.8~0.9	31
0.7~0.8	29
0.6~0.7	26
0.5~0.6	47
<0.5	21
0.0	6

Style	Mean	SD
Animation	0.6672	0.2671
Line Drawing	0.5835	0.2879
Clipart	0.6862	0.2666
Illustration	0.7182	0.2432

Table 7-6. Number of pictures at each accuracy rate level (left), and the mean and standard deviation for different styles (right).

There were specific instances in the Mechanical Turk Study where pictures were tagged incorrectly. Animations vs. static images made no difference in this incorrect tagging. The problems occurred when symptoms were similar enough and the drawing / animation ambiguous enough so that one drawing / animation was tagged to have the meaning of another. For example, an ankle fracture was often tagged as an ankle sprain, sneezing was often tagged as coughing (Figure 7-17), acne was often tagged as chicken pox, and constipation tagged as diarrhea. In another case, the pictures were clear enough but the inadequate medical knowledge of the tagger caused clogged arteries to be tagged as glucose problems and vice versa. There were also possible gender issues, that is, menstrual cramps were tagged as abdominal cramps, potentially by male participants. This suggests that more precise images may highlight the important details of the symptom, but, of course, these images will also have to be tested for their efficacy.

7.3.5 Comprehension of Picture Enhanced Diagnostic Information, Study Design

It would be incredibly difficult to run an experiment diagnosing real patients because each patient's diagnosis will be unique and therefore not comparable. Also, each patient's emotional response to illness will be unique and thus, have an effect on their comprehension of the diagnosis. In addition, as indicated earlier in *Section 7.3.3*, because multiple symptoms are the typical response to a medical problem, we also need an experiment that tests the impact of the presentation of multiple symptoms and causes on their comprehension. I therefore chose a matching experiment and a control measure of medical literacy to examine the viability of OMLA. Native English and non-native English speakers were used to simulate medical literacy problems arising from language barriers.

7.3.5.1 Participants

Ten native and ten non-native English speakers participated in medical information comprehension study. They were the same group that attended the food information comprehension study (*Section 7.2.5*). All of the participants had education at the level of some college or above. In the native language group, ages ranged from 20-65. In the non-native language group, ages ranged from 20-40. Both groups were equally divided in gender. The native languages for the non-native English speakers were Chinese/Mandarin (6), Chinese/Cantonese (1), Russian (1), German (1), and Tamil (1).

7.3.5.2 Assessment of Medical Practices

Before the study, participants were given a questionnaire on their knowledge, attitudes, and experience with medical care:

(1) How much medical terminology did people feel they knew (1/no idea at all ~ 5/very knowledgeable)? The average score for native speakers was 3.5 (standard deviation SD = 0.85), and non-native speakers was 2.25 (SD = 0.79).

(2) How well did people think they could describe their symptoms (1/very poorly ~ 5/very well)? The average score for native speakers was 3.8 (SD = 0.63), and non-native speakers was 2.5 (SD = 0.97).

(3) What did people do when they were feeling sick? Eight out of ten non-native speakers chose to go to the hospital if they felt sick for more than two days. On the contrary, only two native speakers chose to see a doctor, while five took home remedies, three asked family members for help.

(4) What was their experience with various medical care facilities in the past two years? The times people went to a hospital/clinic/medical center or a pharmacist varied greatly, but return visits were almost always for the same problem.

(5) What medications did people keep at home? Nineteen participants keep medications at home. Five of the six Chinese participants mainly had traditional Chinese medications. The most common commercial medicines kept were Tylenol (7), pain reliever (6, e.g. Advil), Aspirin (6), allergy medicines (4), First Aid Kits (4).

(6) How did they find appropriate medications? Eight native speakers tried to figure out what to take on their own, and five of them used a doctor's prescription. Seven non-native speakers found medication through a doctor or family member. Only three non-native speakers said they would search for medicines on their own.

(7) Did they read medical treatment instructions? Whether and how carefully people read instructions varied widely.

Based on participants' self-assessment and responses, their knowledge about illnesses and common American medications was rated (0/no idea at all ~ 5/very knowledgeable). Table 7-7 depicts knowledge level as well English proficiency level (numbers of subjects).

English Proficiency	**Medical Domain Knowledge**			
	a little	**some**	**fair**	**a lot**
fair	2	-	-	-
good	1	4	1	-
excellent	-	-	6	6

Table 7-7. Number of people at each level of English proficiency and medical literacy.

7.3.5.3 Experimental Design and Hypothesis

The diagnostic information comprehension study aimed to reveal the linguistic and cultural competency problem related to medical information comprehension, and evaluate the efficacy of the multimedia (iconic) language support system. The task was to assign 18 medical treatments (from a list of 20) to the illnesses that they were most likely to be used for (one-to-one match).

The matching study was a two (text vs. text + picture) by two (native vs. non-native) within subject factorial design. The dependent variables were accuracy on the matching task and the number of times people viewed a pictorial representation via OMLA. Similar to the food matching study presented in *Section 7.2*, there is a learning effect which can be a confound with the independent variables in the within subject design. However, given the large variance in people's medical literacy, especially among non-native English speakers, a between-subject study with matched pairs was hard to conduct. Therefore, I applied the same scheme as in the food matching study to help control the learning effect in the within-subject design.

The hypotheses for the study were:

H1. People can understand a combination of symptoms and their potential diagnosis significantly better when pictorial representations enhance the text.

H2. Non-native English speakers' performance on the matching task will be significantly more improved than that of the native speakers.

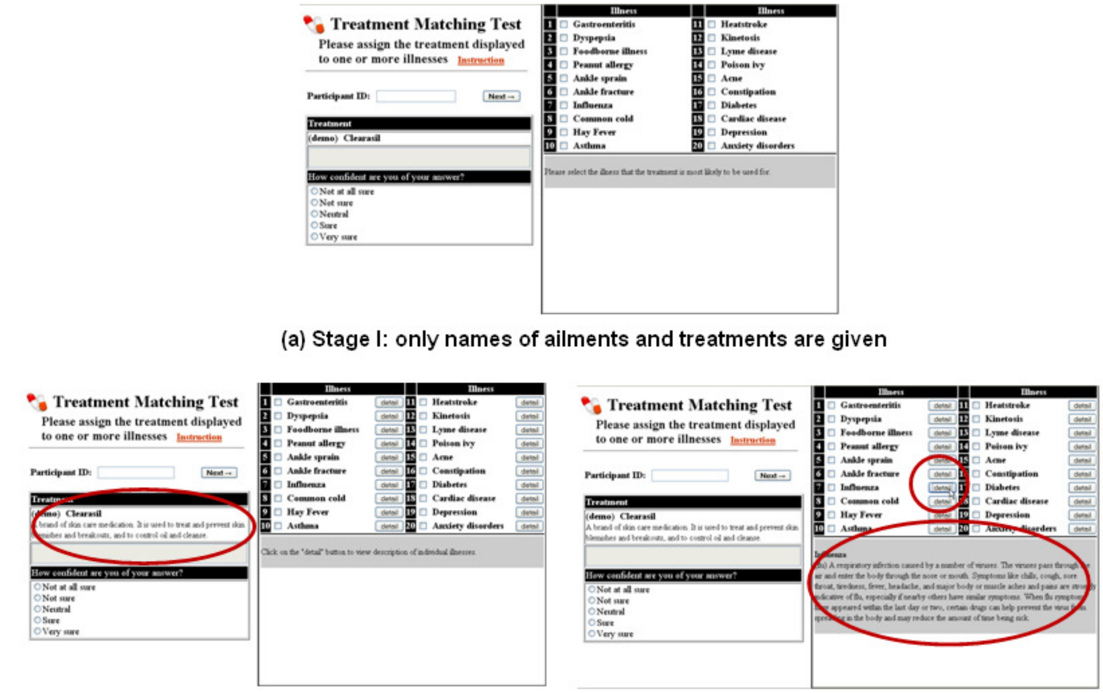
H3. Participants with more medical care domain knowledge (medical literacy) will perform significantly better than those with less knowledge on the matching task (post hoc study).

H4. Non-native English speakers used significantly more pictorial representations than native English speakers.

H5. Even native English speakers can benefit from the pictorial support provided by OMLA.

7.3.5.4 Study Procedure and Interface

In the study, participants were asked to assign eighteen medical treatments to one of twenty illnesses. The study was divided into three stages. In Stage I (Figure 7-18(a)), only the names of the medical treatments and illnesses were given. In Stage II (Figure 7-18(b)(c)), text descriptions of both treatments and illnesses were provided. In Stage III (Figure 7-19 (a)), pictorial representations were made available as a supplement to the text descriptions. This worked as follows: For any highlighted word in the text, participants could click on the word and a pop-up picture or animation would help to explain the meaning of the word (Figure 7-19 (b)(c)(d)).



Treatment Matching Test

Please assign the treatment displayed to one or more illnesses Instruction

Participant ID: Next →

Treatment

(demo) Clearasil

How confident are you of your answer?

○ Not at all sure
○ Not sure
○ Neutral
○ Sure
○ Very sure

	Illness		Illness
1	□ Gastroenteritis	11	□ Heatstroke
2	□ Dyspepsia	12	□ Kinetosis
3	□ Foodborne illness	13	□ Lyme disease
4	□ Peanut allergy	14	□ Poison ivy
5	□ Ankle sprain	15	□ Acne
6	□ Ankle fracture	16	□ Constipation
7	□ Influenza	17	□ Diabetes
8	□ Common cold	18	□ Cardiac disease
9	□ Hay Fever	19	□ Depression
10	□ Asthma	20	□ Anxiety disorders

Please select the illness that the treatment is most likely to be used for.

(a) Stage I: only names of ailments and treatments are given

Treatment Matching Test

Please assign the treatment displayed to one or more illnesses Instruction

Participant ID: Next →

Treatment

(demo) Clearasil

A brand of skin care medication. It is used to treat and prevent skin blemishes and breakouts, and to control oil and cleanse.

How confident are you of your answer?

○ Not at all sure
○ Not sure
○ Neutral
○ Sure
○ Very sure

	Illness			Illness	
1	□ Gastroenteritis	detail	11	□ Heatstroke	detail
2	□ Dyspepsia	detail	12	□ Kinetosis	detail
3	□ Foodborne illness	detail	13	□ Lyme disease	detail
4	□ Peanut allergy	detail	14	□ Poison ivy	detail
5	□ Ankle sprain	detail	15	□ Acne	detail
6	□ Ankle fracture	detail	16	□ Constipation	detail
7	□ Influenza	detail	17	□ Diabetes	detail
8	□ Common cold	detail	18	□ Cardiac disease	detail
9	□ Hay Fever	detail	19	□ Depression	detail
10	□ Asthma	detail	20	□ Anxiety disorders	detail

Click on the "detail" button to view description of individual illnesses.

(b) Stage II: text description of the treatment is given

Treatment Matching Test

Please assign the treatment displayed to one or more illnesses Instruction

Participant ID: Next →

Treatment

(demo) Clearasil

A brand of skin care medication. It is used to treat and prevent skin blemishes and breakouts, and to control oil and cleanse.

How confident are you of your answer?

○ Not at all sure
○ Not sure
○ Neutral
○ Sure
○ Very sure

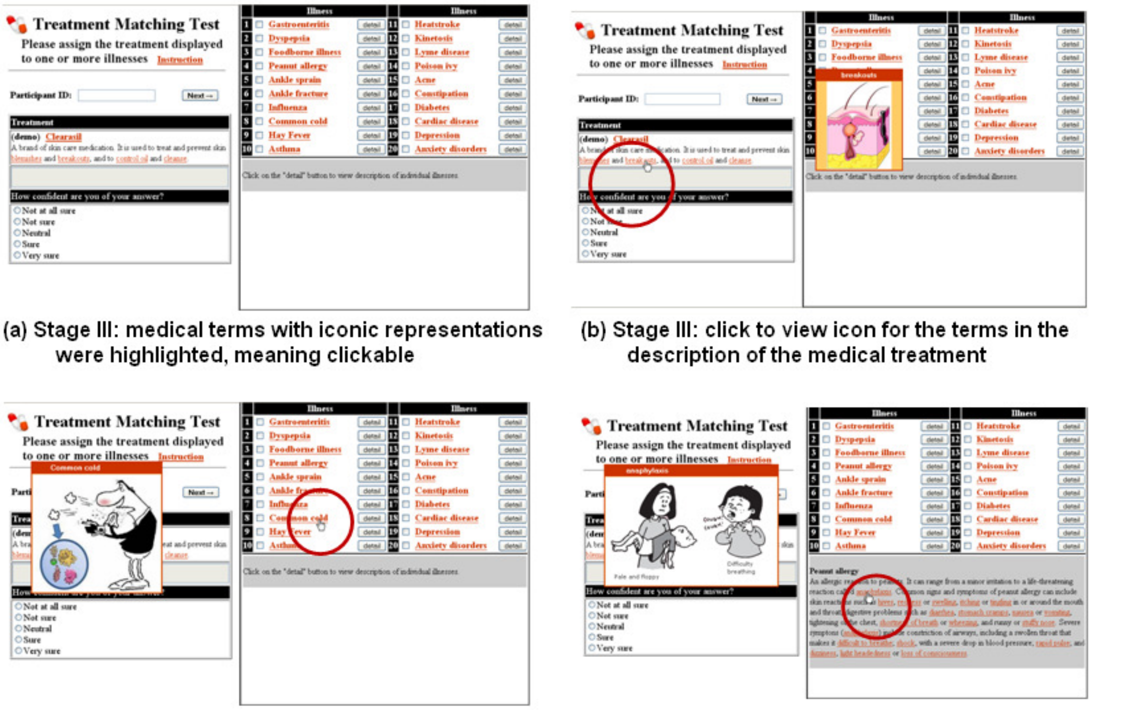
	Illness			Illness	
1	□ Gastroenteritis	detail	11	□ Heatstroke	detail
2	□ Dyspepsia	detail	12	□ Kinetosis	detail
3	□ Foodborne illness	detail	13	□ Lyme disease	detail
4	□ Peanut allergy	detail	14	□ Poison ivy	detail
5	□ Ankle sprain	detail	15	□ Acne	detail
6	□ Ankle fracture	detail	16	□ Constipation	detail
7	□ Influenza	detail	17	□ Diabetes	detail
8	□ Common cold	detail	18	□ Cardiac disease	detail
9	□ Hay Fever	detail	19	□ Depression	detail
10	□ Asthma	detail	20	□ Anxiety disorders	detail

Influenza

(flu) A respiratory infection caused by a number of viruses. The viruses pass through the air and enter the body through the nose or mouth. Symptoms like chills, cough, sore throat, tiredness, fever, headache, and major body or muscle aches and pains are strongly indicative of flu, especially if nearby others have similar symptoms. When flu symptoms have appeared within the last day or two, certain drugs can help prevent the virus from spreading in the body and may reduce the amount of time being sick.

(c) Stage II: text description of the ailment is given

Figure 7-18. Interface of the ailment-treatment matching study, Stage I and II.



Treatment Matching Test

Please assign the treatment displayed to one or more illnesses Instruction

Participant ID: Next →

Treatment

(demo) Clearasil

A brand of skin care medication. It is used to treat and prevent skin blemishes and breakouts, and to control oil and cleanse.

How confident are you of your answer?

○ Not at all sure
○ Not sure
○ Neutral
○ Sure
○ Very sure

		Illness				Illness	
1	□	Gastroenteritis	detail	11	□	Heatstroke	detail
2	□	Dyspepsia	detail	12	□	Kinetosis	detail
3	□	Foodborne illness	detail	13	□	Lyme disease	detail
4	□	Peanut allergy	detail	14	□	Poison ivy	detail
5	□	Ankle sprain	detail	15	□	Acne	detail
6	□	Ankle fracture	detail	16	□	Constipation	detail
7	□	Influenza	detail	17	□	Diabetes	detail
8	□	Common cold	detail	18	□	Cardiac disease	detail
9	□	Hay Fever	detail	19	□	Depression	detail
10	□	Asthma	detail	20	□	Anxiety disorders	detail

Click on the "detail" button to view description of individual illnesses.

Peanut allergy

An allergic reaction to peanuts. It can range from a minor irritation to a life-threatening reaction called anaphylaxis. Common signs and symptoms of peanut allergy can include skin reactions such as hives, redness or swelling, itching or tingling in or around the mouth and throat, digestive problems such as diarrhea, stomach cramps, nausea or vomiting, tightening of the chest, shortness of breath or wheezing, and runny or stuffy nose. Severe symptoms (anaphylaxis) include constriction of airways, including a swollen throat that makes it difficult to breathe, shock, with a severe drop in blood pressure, rapid pulse, and dizziness, light headedness or loss of consciousness.

(a) Stage III: medical terms with iconic representations were highlighted, meaning clickable

(b) Stage III: click to view icon for the terms in the description of the medical treatment

(c) Stage III: click to view animation for the ailment

(d) Stage III: click to view icons for the ailment description

Figure 7-19. Interface of the ailment-treatment matching study, Stage III.

The descriptions of the illnesses came from the NIH website (NIH2009), while the descriptions of treatments came from other online resources such as drugs.com (drugs2009), Wikipedia, etc. The descriptions were 1) tailored to reflect only the major useful facts related to the matching; 2) modified to embed hints like what major treatments should or should not be used for the illness; 3) written to keep the important medical terminologies; 4) changed to add a common “nick name” for the illness; and 5) rewritten so that different words described symptoms in illness category than in the treatment category to prevent participants from using text-matching to generate their answers.

For each treatment selected, the participants indicated their confidence in their assignment. In Stages II and III, people could click on a “detail” button behind each illness to see a textual description of the illness. In Stage III, they could click on the highlighted medical terms to bring up the pop-up drawings / animations. The number of correct assignments at each stage, what details were selected by participants, and what drawings / animations were viewed were automatically recorded by the interface.

The study was piloted with three non-native English speakers. Based on feedback the following modifications were made to the initial design:

(1) Removed medications like aspirin, Advil, and Tylenol, which were included in the initial treatment list, because they could be applied to several illnesses and their usage varied according to personal experience.

(2) Fit all the content onto one page. In the initial design, all illness descriptions were displayed at the same time, requiring a lot of scrolling which made the matching a difficult recall task.

(3) Changed to only one possible match. Originally, participants were asked to assign treatments to all the possible illnesses that could be helped by a particular treatment. However, results varied greatly due to personal experience. For treatments that are applied equally frequently to multiple ailments, assigning any of them is considered correct. For example, Prozac has been approved for the treatment of major depression, obsessive-compulsive disorder, and panic disorder.

(4) Added feedback on performance for each stage. In the pilot studies, participants had no idea how well they performed, and thus showed overconfidence and were less careful in checking their work. Thus, interface was modified so that it told them how many mistakes they made at the completion of each stage.

7.3.6 Results and Analysis

7.3.6.1 Comparison of Performance of Native and Non-Native Speakers at Each Stage

The dependent variables (accuracy and number of pictures viewed) were tested for homogeneity of variance using the Bartlett and Flinger-Killeen tests. Both variables were found to not have distribution problems.

Figure 7-20 shows the overall matching performance across all three stages. The improvement across each of the three stages was significant (F(2, 54) = 105.549, $p < 0.001$, $\eta^2 = 0.629$). The difference between correct matches and confident correct matches in Stage I indicated that given only the names of medical treatments and illnesses, about one quarter of the correct assignments were guesses.

Performance between the native and non-native speakers was significantly different (F(1,54) = 49.588, $p < 0.001$, $\eta^2 = 0.148$). The interaction between language group and stage was also significantly different (F(2,54) = 10.376, $p < 0.001$, $\eta^2 = 0.062$), suggesting that native and non-native speakers performed differently at each stage.

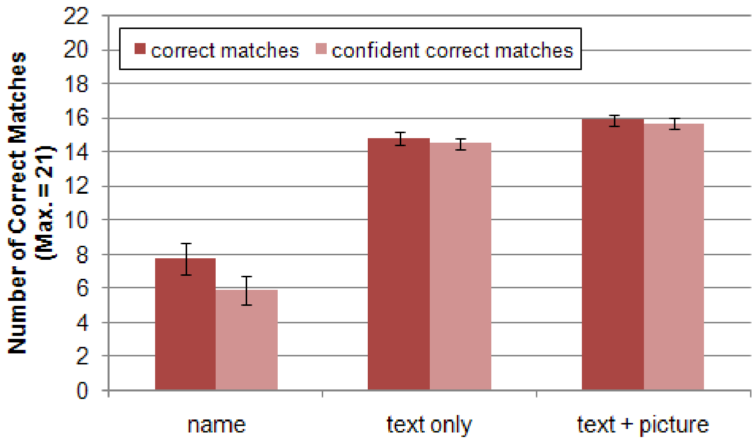


Figure 7-20. Comparison of number of correct matches at each stage (the light red columns show the correct matches which people were sure of).

Figure 7-21 shows the number of correct matches (eliminating those that were guesses) and the number of confident selections ("sure" or "very sure") by the two groups in Stage I and II. At both stages (Stage I: F(1,18) = 53.894, $p < 0.01$, $\eta^2 = 0.7496$; Stage II: F(1,18) = 10.286, $p < 0.01$, $\eta^2 = 0.3636$) native speakers outperformed non-native speakers. In Stage I given names only, native speakers, in general, know more about common over-the-counter medicines, as is also shown by the confidence rating. However, the difference between the confident matches and the correct matches implies that people's prior knowledge is possibly wrong. In Stage II, the difference between the language groups decreased greatly. Text descriptions did provide useful but not always sufficient information.

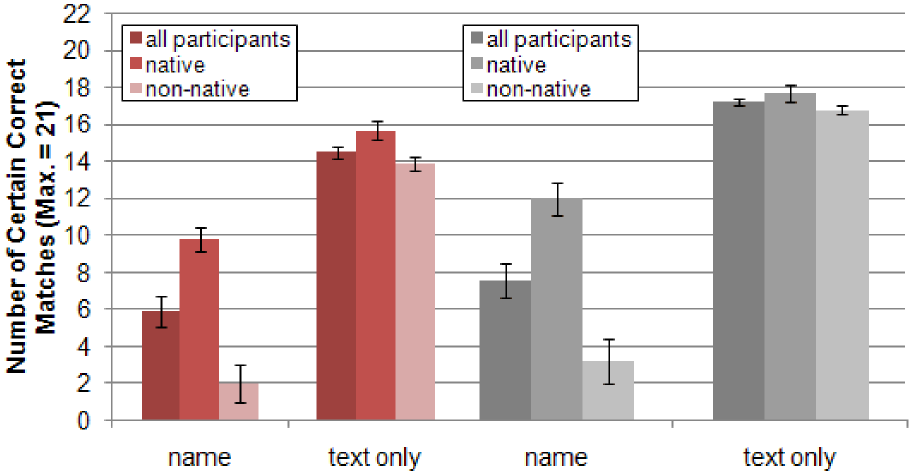


Figure 7-21. Increase of accuracy (left, red) and confidence (right, grey) in Stage I and II.

The situation in Stage III was more complicated: 1) in some cases, participants did not view any drawings/animations if they felt sure of their choice or felt that they understood the entire text; 2) participants viewed icons/animations to determine the meaning of the words; and 3) participants clicked on the drawings/animations just for fun ("I was curious how you depicted this.") Figure 7-22 shows the average number of picture clicks of the two language groups. Non-native speakers viewed significantly more drawings/animations than native speakers (F(1,18) = 7.6709, p = 0.013, η^2 = 0.8847). Figure 7-23 shows the Stage III number of correct matches with and without the assistance of icons/animations.

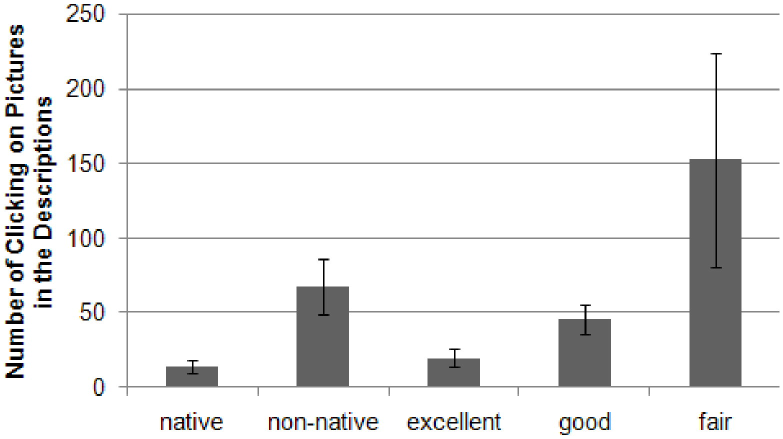


Figure 7-22. Number of times people with different English proficiency viewed the pictures in Stage III.

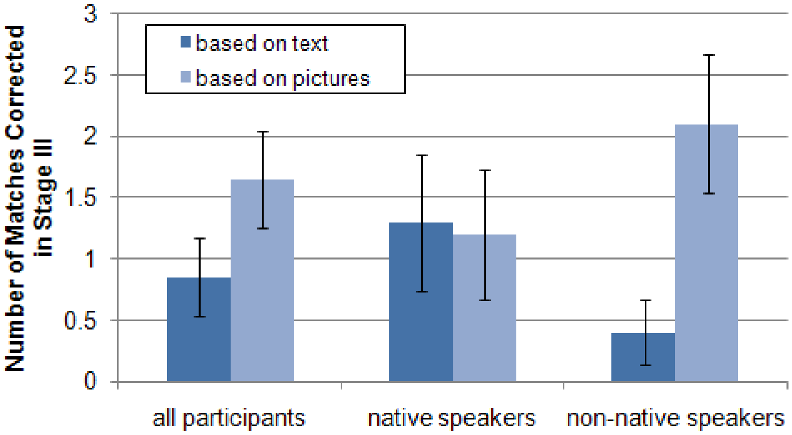


Figure 7-23. Number of correct matches in Stage III based entirely on text vs. with the help of icons/animations, for native and non-native English speakers.

The improvement from using pictures for non-native speakers was significantly greater than that for native speakers (F(1,18) = 3.7556, p = 0.034, η^2 = 0.790). We examined the performance of the individual participants and found that the results from the non-native speakers who speak very fluent English and have large English vocabularies were similar to the results from the native speakers. This suggests that English proficiency might be a more accurate way to classify than simple native/non-native speakers. What is more, there might be other influential factors such as amount of medical domain knowledge. Therefore, we did some post-hoc tests on these factors but our N was too low to get more than suggestive results.

7.3.6.2 Post-hoc Test: Influence of English Proficiency and Medical Knowledge

In the post-hoc tests, we looked into the influence of English proficiency (Figure 7-24 left) and medical domain knowledge (Figure 7-24 right) as well as their interaction in:

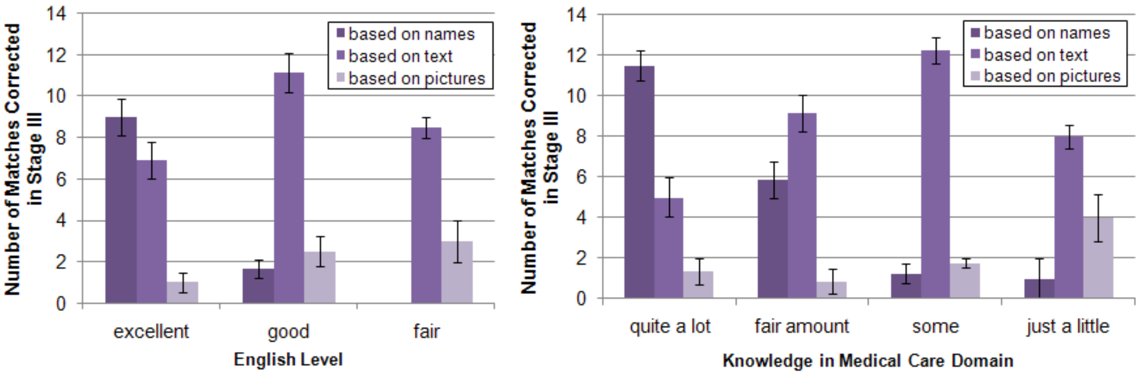


Figure 7-24. Number of correct matches based on names (Stage I), entirely on text (Stage II and III) vs. with the help of pictures (Stage III) for different English levels (left) and medical knowledge (right).

(1) The number of correct matches in Stage I. The impact of both English proficiency (F(1,16) = 86.513, $p < 0.001$, $\eta^2 = 0.657$), domain knowledge (F(1,16) = 16.502, $p < 0.001$, $\eta^2 = 0.125$), as well as their interaction (F(1,16) = 12.582, $p < 0.001$, $\eta^2 = 0.096$) was significant. Not surprisingly, a person's English proficiency was highly correlated with a person's medical literacy since the matching test involved American medicine terms.

(2) The number of correct matches (which were wrong in Stage I) rose by using the text descriptions only in Stage II. The results suggest that participants can figure out answers from text alone based on the high levels of English proficiency and domain knowledge (F(1,16) = 11.422, $p < 0.001$, $\eta^2 = 0.305$), although English proficiency, by itself, still showed significant impact (F(1,16) = 6.294, $p = 0.023$, $\eta^2 = 0.168$). If a person did not have a large enough vocabulary to understand the description of the illnesses, of course, it will be hard for them to find the right treatment. On the other hand, without pre-learned knowledge about an illness, even if one is able to understand individual words in the description, he or she might still not be able to integrate all the symptoms into a unique diagnosis. An example is "peanut allergy." None of the Chinese participants knew that peanuts can cause a life-threatening allergy reaction because this situation is rare in China and thus did not consider a treatment

for severe allergic shock. At Stage II, nine out of ten non-native speakers made a mistake on the “peanut allergy” assignment, while only one native speaker had it wrong.

(3) The number of correct matches rose with the use of drawings/animations in Stage III. Only English proficiency significantly affected how much the drawings/animations helped with a participant’s comprehension ($F(1,16) = 9.591$, $p < 0.001$, $\eta^2 = 0.355$). This result suggests that for people who are not yet fluent in English, icons/animations are a helpful support. For example, consider the medical term “urushiol” which is the chemical oil in poison ivy that causes an allergic skin response. Many people failed to locate the word in the text descriptions, but once they saw the picture of poison ivy that represented “urushiol,” they immediately picked up on the connection.

7.3.6.3 Results by Medical Treatments

Table 7-8 lists all of the medical treatment – illness assignments that everyone in each language group got right. The improvement in Stage III over Stage II is highlighted in bold font. From an examination of our data, it is clear that all improvements arose from participants viewing the icons/animations. The final results also aligned with the results from the tagging study. The illnesses that matching study participants made the most mistakes on (in both Stage II and III) were gastroenteritis (stomach flu), influenza (flu), and Lyme disease. This suggests that these illnesses were hard to describe visually AND verbally.

	Native Speakers		**Non-Native Speakers**
Stage II	Calamine lotion Clearasil Ex-lax Gatorade Insulin	ORIF ProAir HFA Prozac Statin	Calamine lotion ORIF Prozac RICE Protocol Statin
Stage III	Clearasil (poison ivy) **EpiPen** (peanut allergy) Gatorade (heatstroke, gastroenteritis, foodborne illness) Insulin (diabetes) **Pepto Bismol** (dyspepsia) Prozac (depression, anxiety) Statine (cardiac disease)	**Dramamine** (kinetosis) Ex-lax (constipation) ORIF (ankle fracture) ProAir HFA (asthma) **RICE Protocol** (ankle sprain) **Zyrtec** (hay fever)	Calamine lotion **Clearasil** **Dramamine** ORIF **ProAir HFA** Prozac RICE Protocol Statin **Tecnu**

Table 7-8. Lists of successful matches by all people in each group at Stage II and III.

7.3.7 Discussion

All of the hypotheses presented were found to be significant. The pictorial representations were particularly helpful in discrimination cases were a particular symptom eliminated similar diagnoses. This was true for both native and non-native English speakers as well as for low and high medical literacy personnel. Thus, it can be concluded that having pictorial representations available for all diagnoses is a useful addition. Observations were noted during the study, and useful feedback from the participants was also collected.

(1) The animations were too long (approximately five seconds each). Our participants did not have the patience to watch the entire animation and frequently quit after viewing only five frames. It raises the question of how to retain attention, especially in a doctor's office with multiple distractions.

(2) Many non-native speakers (especially those with good to fair English) used animations as a quick check on their understanding of an illness. Matching twenty illnesses can be a memory load for non-native speakers. The animations provided information on the type of disease (skin, digestion, or heart problem) at a glance causing them to bypass reading the text. It suggests that animations can facilitate cued recall.

(3) The animations helped fix people's misunderstanding of the text. For example, several people initially mistook “heatstroke” for “heart-stroke.” However, sometimes animations or pictures in general can interfere with a person's comprehension. For instance, the animations for stomach flu, food poisoning, and upset stomach are not distinctive enough so that two native-speakers got confused after looking at the pictures.

(4) People made the assignments by “keyword matching.” However, if the same symptom was described using different phrases in the treatment and illness textual descriptions respectively, this strategy failed. When this happened, people looked for similarity in the animations.

(5) Many native English speakers matched medicines based on superstition/prior beliefs from life experience and did not bother to check the pictorial representations. For example, one native speaker who thought “Tecnu” is used for “Acne” did not view any of the pictures.

(6) People do not always read carefully. For example, we wrote in the description for food poisoning that “The main treatment is putting fluids back in the body,” but people still selected antibiotics indicating that people have trouble following long text descriptions of health information.

Past experience affects a person's knowledge of medical care. The three participants above age 40 knew more about illnesses and had used or heard of various treatments. Younger Americans, in contrast, mentioned that they had heard of some of the medications on commercials, but could not remember what they were for. Non-native English speakers who had little experience with Western medications only knew medicines from their home countries. Much of their knowledge of illnesses and western medications came from American commercials and movies.

7.3.8 Conclusions

Section 7.3 introduces the use of the Online Multimedia system to support medical care information understanding across language barriers. Two methods for measuring the precision and comprehension of picture-enhanced diagnostic health information were presented. The first was an online tagging study to verify the comprehensibility of the chosen pictorial representations for medical terminologies via Amazon Mechanical Turk. Although little was known about the subject population in this study, the large N coupled with the embedded checks for response bias made it an efficient method for measuring the precision of large scale pictorial representations required in the OMLA system. The style of the pictorial stimuli did not show a significant impact. However, the detail in the content portrayed in the pictures affected the precision and comprehension. The second study was a treatment-illness matching experiment that compared native English to non-native English speakers' performance with and without the assistance of the pictorial representations. Not surprisingly, the findings suggested that language ability and medical literacy had a significant effect on people's understanding of medical care information. All of the hypotheses were confirmed. OMLA was especially helpful for those with low English proficiency and also for native speakers possessing weak medical knowledge. These results suggest that a multimedia language support system is a useful and viable improvement to diagnostic procedures and health care related information comprehension.

7.3.9 Multimedia Communication Support Applications, Summary

Chapter 7 has presented the design, development, and evaluation of the Online Multimedia Language Assistant (OMLA) for people with language disabilities or barriers. The main function of the language assistant is an online popup dictionary that allows users to click on an unfamiliar word on any webpage they are browsing to instantly look up multimedia representations of the selected word (if any) stored in the backend multimedia-enhanced lexical library. Users can view and customize the dictionary and vocabulary by either text search (if they can spell out the word) or directly navigating into the library. Users can also upload their own pictures, videos, and audio clips to the multimedia library to create a personalized vocabulary that fits their daily topics.

The Multimedia Augmented Online Language Assistant has the following features distinctive from other augmentative and alternative communication systems.

1) The language assistant targets normal Internet information accessibility and understanding for people with language impairment.

2) It extends the convenient visual representations (stylized icons and animations) in assistive technologies to a much wider scope, adding web images, videos, and non-speech audio clips.

3) The popup multimedia dictionary of the language assistant is implemented as a browser plug-in enabling users to manipulate web content directly. Unlike other pictorial dictionaries websites for cross-language translation, the popup dictionary minimizes the operations required to lookup an unknown word and avoids the loss of context due to switching between webpages for users who have language impairment and maybe other cognitive/motor disabilities due to a stroke, brain tumor or head injury – major cause of aphasia.

4) Different from most of the assistive devices and pictorial dictionaries in which the visual representations are created or selected by the designers, the language assistant utilizes the abundant web images and provides adaptable mechanisms for the users to contribute to and personalize the multimedia library, increasing the scalability of the language assistant.

Two studies in real information comprehension scenarios (food and medical care) were conducted to evaluate the effectiveness of OMLA in enhancing people’s understanding of unfamiliar terminologies. It has been shown that non-native English speakers can benefit from the use of multimedia language support extensively. Even native speakers received useful assistance from the multimedia representations when there is a lack of domain knowledge. These findings justify the viability of the multimedia vocabulary developed throughout my thesis research in the area of Augmentative and Alternative Communication.

CHAPER 8 CONCLUSIONS AND FUTURE WORK

8.1 Summary of Methodologies, Findings and Design Guidelines

In this dissertation, I have presented my Ph.D. work on multimedia approaches to facilitating communication beyond words. Upon identifying limitations in current Augmentative and Alternative Communication techniques for people with language disabilities and barriers, I conducted research that provides the theoretical grounds for utilizing semantically organized multimedia data as information carriers, which greatly extends the scope of current AAC vocabularies in terms of size, media richness, and usability.

Over 5,000 web images, over 3,000 icons and animations, 50 videos and more than 300 environmental sounds were collected and evaluated as representations for the most frequently used nouns, verbs, adjectives, and adverbs. Ten studies (two for static pictures, two for dynamic pictures, two for sounds, one usability study and three user studies for the Online Multimedia Language Assistant) were carried out to gather information that answers the following research questions:

(1) What kinds of non-verbal representations are effective, inexpensive, and easy to scale?

Web images which are numerous, varied, and cheap are good enough representations for nouns. Videos and environmental sounds with increasing availability as the Internet expands have demonstrated their success in communicating a large subset of common concepts.

(2) How can information be encoded with non-verbal representations so that message can be unambiguously conveyed?

Review of literature and empirical knowledge suggest that non-verbal representations should be linked to concepts instead of words which could be polysemous. Unfortunately, the representation-to-word association is what conventionally used in AAC vocabularies. A new algorithm was presented and tested that can automatically reconstruct (with overall 80% accuracy) the connection between concepts and images/sounds based on their labels. Given the availability of huge image/sound label datasets over the Internet, building a large multimedia-enhanced semantic network that can satisfy people's communication needs becomes feasible.

(3) How can the information and illustrations be organized so that they can be efficiently accessed to support timely communication?

Categories, scenarios, frequency and some other organization schemes that have been employed by AAC vocabulary are found to be inefficient in navigation and word finding in many cases. Theories and models on mental lexicon suggests that a vocabulary structure that simulates how concepts and words are stored and activated in the human brain is more effective for timely retrieval. Representations for different concepts should be linked according to various semantic relationships, so that related representations for a given context are more useful for information comprehension and expression.

(4) Can AAC enhance the comprehension and usage of written information online, and how?

Yes, in addition to sentence composition, AAC vocabularies can serve as a non-verbal dictionary which can provide visual/auditory clues for the unknown concepts and messages. Web browser extensions are a valid tool for this purpose since they allow the access of multimedia dictionary without switching out of the web browsing context. Studies showed that non-native English speakers can greatly benefit from the Online Multimedia Language Assistant. Even native speakers received substantial help when trying to understand written information in domains that they have limited knowledge of.

Here, I would like to highlight several aspects of my Ph.D. research that distinguish it from previous work. According to my knowledge, this research is the first that explores the use of videos and environmental sounds as representations for Augmentative and Alternative Communication. It is one of the few researches that compared more than two stimuli in evaluating efficacy of representations in both individual naming and comprehension in sentence context. The size of the tested vocabulary and representation inventory is the biggest known, especially with videos and environmental sounds. One of the studies was conducted with 50 people with aphasia, which is very rare in research with this population given the limited accessibility to the target population. Another study that was carried out online involved over 2,000 anonymous subjects from 46 countries, which is also a striking number, especially given the small amount of money we spent on the study. A new experiment method, i.e. the use of Amazon Mechanical Turk, is explored, and strategies to increase the control of result quality are proposed since little is known about the participants and their behaviors during the study. These strategies can inspire other researchers who wish to make use of this platform as well.

According to the literature review in the field of AAC, design rules have only been given to graphical icons. Based on the findings and observations in my studies, I propose the following guidelines for creating/selecting icons, images, animations, videos, and sounds as representations for common concepts.

(1) Representations should have good enough quality to present the visual/auditory details, especially those obtained from the Internet.

(2) Representations should eliminate unnecessary details to avoid distraction, such as texts in the background and facial expressions of the characters when the message is unrelated to emotions and feelings.

(3) Representations should highlight the intended message if possible.

(4) Representations should take advantage of commonly accepted knowledge, such as gestures/poses, universal symbols, and distinctive audio patterns and sound events.

(5) Representations should avoid the choice of content based on unique knowledge or experience from a small subset of people.

(6) Representations should avoid using indirectly encoded messages or ambiguous elements, such as non-intuitive symbols or acoustically similar sounds.

(7) Representations should choose the most unique aspect of the concept among all the possibilities, such as different sounds a tool can make or different view points of an object.

(8) Representations should be careful of the use of examples from a distinctive sub-category to convey a more generic concept. Prototype theory shows that people's attention may be shifted to the wrong level of details.

(9) Representations should be sensitive to age differences, in terms of age-related degenerations and variance in familiarity due to life experiences.

(10)Representations should be sensitive to cultural differences, which may affect the choice of symbols, gestures, and sound scenes.

(11)Representations should be visually/sonically pleasing if applicable. The goal is the facilitate communication, not to disturb the users or cause social embarrassment. In addition, anecdotes show that people are more likely and more willing to use a vocabulary that they find pleasant and fun.

These guidelines are helpful for directing future development of any modality of concept representations for Augmentative and Alternative Communication.

8.2 Future Work

Future research extends in three directions: representation design and validation, vocabulary adaptability and adaptiveness, and real life applications.

Representation design and validation: my thesis research has proven that concrete nouns and concrete verbs can be successfully conveyed through pictures, animations, and videos. However, how to depict abstract concepts and attributes (adjectives and adverbs) effectively through visual and auditory representations remains unknown. As abstract nouns/verbs, adjectives and adverbs take up a big proportion of a person's working vocabulary and are critical to the richness of communication, ways to convey these parts of speech in AAC should be studied. Based on the knowledge established in my thesis research, I would like to explore the means and guidelines to create good carriers for abstract concepts and attributes.

Furthermore, up to this point, each representation modality was tested individually. The communicative power of a combination of representations across modes is under-investigated. For example, will a picture and an environmental sound for the same concept given together be better perceived than given separately? This is an interesting but hard problem. First of all, not all the concepts can be equally well evoked visually and auditorily. Second, the best pictorial representation may not match the most distinctive sound. How visual and auditory perception interfere with each other, how different aspects are registered onto the same concepts, how pictures and sounds should compromise to achieve the best illustration performance are all interesting questions to explore.

The evaluations of the multimedia language were primarily carried out with the unimpaired user population, i.e. people with language barriers. However, individuals with language disabilities may perform differently since other parts of their cognition and/or motor functionalities may have been affected as well. I would like to test the multimedia-enhanced semantic network with the language impaired population in order to draw more general conclusions about its effectiveness.

Vocabulary adaptability and adaptiveness: as discussed in the background work chapter, the fringe vocabulary (in contrast to the core vocabulary) varies greatly from person to person. What words are known, how frequently they are used, and how words and concepts are linked change as one's age, gender, education, social status, profession and hobby differ. It suggests that any AAC system should provide adaptable features that allow users to customize the size, content, and organization of the vocabulary. Furthermore, a smart vocabulary should be able to take user input and usage history to adapt its structure and display automatically, saving the time and efforts for the people who already have communication difficulties. How to balance between the adaptable and adaptive components is another important issue. I would like to work on approaches to promote the efficacy of the multimedia language by integrating adaptable and adaptive features.

Real life Applications: the communication support system based on the multimedia language has shown its success in improving information comprehension across language barriers in various matching tasks. In real life, a lot of areas that involve intensive information exchange demand such communication assistance. Diagnostic communication in clinical encounters is one of the examples. I would like to design new specialized systems that utilize the rich multimedia language to facilitate oral and/or written communication in actual settings.

Overall, this thesis demonstrates that the use of multimedia as a communication aid can be a much more effective tool than current single picture approaches.

References

- A.D.A.M. Medical Encyclopedia. A.D.A.M., Inc. http://www.adam.com/. 2009.
- Ageless Project. http://jenett.org/ageless/. 2008.
- Aitchison, J. *Words in the mind: an introduction to the mental lexicon*. Basil Blackwell, Oxford. 2004.
- Altarriba, J., Bauer, L.M., and Benvenuto, C. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior research methods, instruments, & computers*: a journal of the Psychonomic Society. 1999.
- Amazon Mechanical Turk. https://www.mturk.com/ 2009.
- American Heritage Dictionary of the English Language, Fourth Edition. Houghton Mifflin Company. 2006.
- American Speech-Language-Hearing Association. ASHA Special Interest Division 12, *Augmentative and Alternative Communication* (AAC). http://www.asha.org/docs/html/PS2005-00113.html. 2010.
- American Speech-Language-Hearing Association. Augmentative and Alternative Communication Decisions. http://www.asha.org/public/speech/disorders/CommunicationDecisions.htm . March 2010.
- Anderson, J. and Fisher, B. The Myth of Persistence of Vision Revisited. *Journal of Film and Video.* 45(1): 3-12. 1993.
- AP Images. http://apimages.ap.org/. March 2010.
- Ariew, R. A template to generate hypertext and hypermedia reading materials: It's design and associate research findings. *The Reading Matrix.* 6(3): 195-209. 2006.
- Ascher, S and Pincus, E. *The Filmmaker's Handbook*. London: Penguin. 1999.
- Association of Community Organizations for Reform Now. Speaking the Language of Care: Language Barriers to Hospital Access in America's Cities. 2003.
- Austin, P., Matlack, R., Dunn, K., Kosler, C., and Brown, C. Discharge Instructions: Do Illustrations Help Our Patients Understand Them? *Ann Emerg Med*. 25: 317-320. 1995.
- Baecker, R., Small, I., and Mander, R. Bringing icons to life. In *Proc. of CHI 1991*, ACM Press 1-6. 1991.
- Baker, B. Using images to generate speech. Byte, 11;160-168. 1986.
- Ballas, J.A. Common factors in the identification of an assortment of brief everyday sounds. J. of Experimental Psychology, 19(2):250–267, 1993.
- Ballas, J.A. and Howard, J.H. Interpreting the language of environmental sounds. Environ. Behav., 1(9(1)):91–114, 1987.
- Ballas, J.A. and Mullins, R. Effects of context on the identification of everyday sounds. Human performance, 4:199–219, 1991.
- Ballas, J.A. and Sliwinsky, M.J. Causal uncertainty in the identification of environmental sounds. Tech Report ONR-86-1, Office of Naval Research, Department of Psychology, Georgetown University, Washington, D. C., November 1986.

- Banerjee, S. and Pedersen, T. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*. 2002.
- BBC Sound Effects Library. http://www.sound-ideas.com/bbc.html. 2009.
- Begault, D., Wenzel, E., Shrum, R., and Miller, Joel. A Virtual Audio Guidance and Alert System for Commercial Aircraft Operations. *ICAD'96* 1996.
- Benton, A.L. and Hamsher, K. *Multilingual Aphasia Examination*. University of Iowa. 1977.
- Beukelman, D.R, McGinnis, J., and Morrow, D. Vocabulary Selection in Augmentative and Alternative Communication. Augmentative and Alternative Communication. 7(3) 171-185. 1991.
- Beukelman, D.R. and Mirenda, P. Augmentative and Alternative Communication: Management of Severe Communication Disorders Children and Adults. Baltimore: Paul H. Brookes Publishing Co. 1992.
- Beukelman, D.R. & Mirenda, P. Augmentative and Alternative Communication. Supporting Children and Adults with Complex Communication needs. 3rd edition. Baltimore: Paul H. Brookes Publishing Co. 2005.
- Bird, H, Franklin, S., and Howard, D. Age of Acquisition and Imageability Ratings for a Large Set of Words, Including Verbs and Function Words. *Behavior Research Methods, Instruments, & Computers*. 33(1), 73-79. 2001.
- Bjork, E.A. The Perceived Quality of Natural Sounds. *Acustica*, Research Notes. 57, 185-188, 1985.
- Blackstone, S. Clinical news: Visual scene displays. Augmentative Communication news, 16(2), 1-8. 2004.
- Blank, M, Gorelick, L, Shechtman, E, Irani, M, & Basri, R. Actions as space-time shapes. *Proc. Of ICCV*, 2005.
- Blattner, M., Sumikawa, D., and Greenberg, R. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*. 4(1). 1989.
- Bleser, R.D. and Kauschke, C. Acquisition and Loss of Nouns and Verbs: Parallel or Divergent Patterns? *Journal of Neurolinghistics*. 16, 213-229. 2003.
- Blissymbol Communication (UK). http://www.blissymbols.co.uk/. March 2010.
- Borman, A., Mihalcea, R., Tarau, P. PicNet: Pictorial Representations for Illustrated Semantic Networks, *In Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors*, Stanford, CA, March 2005.
- Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. Adding Dense, Weighted Connections to WordNet. *Proceedings of the Thirds International WordNet Conference*. 2006.
- 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.
- Brach, C. and Fraser, I. Can Cultural Competency Reduce Racial and Ethnic Health Disparities? A Review and Conceptual Model. *Medical Care Research and Review*. 57: 181 (2000).

- Brazil, E. and Fernstrom, M. Empirically based auditory display design. In Sound and Music Computing Conference (SMC 2009), pages 7–12, Porto, Portugal, 2009.
- Brewster, S. Using nonspeech sounds to provide navigation cues. *ACM Transaction on Computer-Human Interactions*. 5(3), pp. 224-259. ACM Press. 1998.
- Carello, C., Anderson, K.L., and Kunkler-Peck. Perception of Object Length by Sound. *Psychological Science*. 9(3):211-214. 1998.
- Carlson, F. A format for selecting vocabulary for the nonspeaking child. Language. *Speech, and Hearing Services in Schools*, 12, 140-145. 1981.
- Carney, R.N. and Levin J.R. Pictorial Illustrations Still Improve Students' Learning From Text. *Educational Psychology Review*. 14(1), pp. 5-26. 2002.
- Carroll, D. Psychology of Language, third edition, Brooks/Cole Publishing Company, p.334-346,1999.
- Clarke, S., Bellmann, A., De Ribaupierre, F., and Assal, G. Non-verbal auditory recognition in normal subjects and brain-damaged patients: Evidence for parallel processing. *Neuropsychologia*. 34 (6), 587-603. 1996.
- Clifton, Z. *AphasiaFox: A Mozilla Firefox Extension Opening up the Internet for People Suffering from Aphasia. Independent Work Report (unpublished).* Princeton University. 2007.
- Clipart.com, 10 Million Downloadable Images by Subscription! http://www.clipart.com/en/. March 2010.
- Code, C., Hemsley, G., Herrmann, M. The emotional impact of aphasia. *Seminars of Speech and Language*. 20(1): 19-31.1999.
- Coleman, G.W. The Sonic Mapping Tool. PhD thesis, University of Dundee, August 2008.
- Collins, A.M. & Loftus, E. F. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.
- Collins, A. M. & Quillian, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247. 1969.
- Corel Image Database. http://wang.ist.psu.edu/docs/related/. 2007.
- Danielsson and Jonsson. Pictures as Language. *Language and Visualization*. 2001.
- Davidoff, J. and Masterson, J. The Development of Picture Naming: Differences between Verbs and Nouns. *Journal of Neurolinguistics*. 9(2), 69-83. 1996.
- Davidson, B., Howe, T., Worrall, L., Hickson, L., and Togher, L. Social participation for older people with aphasia: the impact of communication disability on friendships. *Top Stroke Rehabil*. 15(4):325-40. 2008.
- Delp, C. and Jones, J. Communicating Information to Patients: The Use of Cartoon Illustrations to Improve Comprehension of Instructions. *Acad Emerg Med*. 3: 264-270. 1996.
- 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.
- Dick, F., Bussiere, J., and Saygm, A. The Effects of Linguistic Mediation on the Identification of Environmental Sounds. Newsletter of the Center for Research in Language. 14 (3). 2002.
- Drager, K., Light, J., Carlson, R., D'Silva, K., Larsson, B., Pitkin, L., & Stopper, G. Learning of dynamic display AAC technologies by typically developing 3-year-olds: Effect of different

layouts and menu approaches. *Journal of Speech, Language, and Hearing Research, 47,*1133-1148. 2004.

- Drugs.com. http://www.drugs.com/. 2009.
- Druks, J. and Masterson, J. *An Object and Action Naming Battery*, Psychology Press, London, 2000.
- DynaSyms. http://www.poppinandcompany.com/dynasyms/index.shtml. March 2010.
- DynaVox Mayer-Johnson. http://www.dynavoxtech.com/. March 2010.
- DynaVox System LLC. Selecting and Organizing Vocabulary for AAC Users. 2010.
- Elliecards, early learning images. http://www.elliecards.com/product.html. March 2010.
- Engelter, S.T, Gostynski, M., Papa, S., et al. Epidemiology of aphasia attributable to first ischemic stroke: incidence, severity, fluency, etiology, and thrombolysis. *Stroke*; 37:1379-1384. 2006.
- Eysenck, H. J. A critical and experimental study of color preferences. *American Journal of Psycholog. 72,* 487-502. 1959.
- Fallon, K., Light, J., & Achenbach, A. The semantic organization patterns of young children: Implications for augmentative and alternative communication. *Augmentative and Alternative Communication,* 19, 74-85. 2003.
- Faubert, J. Visual Perception and Aging. *Canadian Journal of Experimental Psychology*. 56(3), pp. 164-176. 2002.
- Fellbaum, C. *WordNet: An electronic database*. MIT Press, Cambridge, MA. 1998.
- Fillmore, C.J. Frame Semantics. *Cognitive linguistics: basic readings*, 1982.
- FindSounds. http://www.findsounds.com/. 2008.
- Flickr, photo sharing. http://www.flickr.com/. March 2010.
- Flippo, K. F., Inge, K. J., and Barcus, J. M. *Assistive Technology: A Resource For School, Work, and Community*. Baltimore: Paul H. Brookes Publishing Co. 1995.
- Fotosearch.com, Stock Photography and Stock Footage. http://www.fotosearch.com/. 2010.
- FrameNet. http://framenet.icsi.berkeley.edu/. March 2010.
- Free Online Dictionary, the. http://www.thefreedictionary.com/. March 2010.
- Freed, D.J. Auditory Correlates of Perceived Mallet Hardness for a set of Recorded Percussive Sound Events. *Journal of the Acoustical Society of America*. 87(1):311-322. 1990.
- Freesound Project. http://www.freesound.org/. 2008.
- Fuller, D.R., Lloyd, L.L., and Schlosser, R.W. Further Development of an Augmentative and Alternative Communication Symbol Taxonomy. *Augmentative and Alternative Communication*. 8(1) 67-74. 1992.
- Garzonis, S., Jones, S., Jay, T., and O'Neill, E. Auditory Icon and Earcon Mobile Service Notifications: Intuitiveness, Learnability, Memorability and Preferences. In Proc. *CHI'09*. pp. 1513-1522. 2009.
- Gaver, W. Everyday listening and auditory icons. Doctoral Dissertation, University of California, San Diego. 1988.
- Gaver, W. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction*. 4. 1989.

- Gaver, W. What in the World Do We Hear? An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, 5 (1): 1-29. 1993.
- Gaver, W., Smith, R., and O'Shea, T. Effective Sounds in Complex Systems: The ARKola Simulation. In Proc. *CHI'91*. pp. 85-90. ACM Press, 1991.Handel, S. *Listening: An introduction to the perception of auditory events*. Cambridge, MA. MIT Press. 1989.
- Gazmararian, J., Baker, D., Williams, M., Parker, R., Scott, T., Green, D., Fehrenbach, S., Ren, J., and Koplan, J. Health Literacy Among Medicare Enrollees in a Managed Care Organization. *Journal of American Medical Association*. 281(6): 545-551. 1999.
- General Text Matcher. http://nlp.cs.nyu.edu/GTM/. 2006.
- Gentner, D. On Relational Meaning: The Acquisition of Verb Meaning. *Child Development*. 49, 988-998. 1978.
- Gentner, D. Why Nouns are Learned before Verbs: Linguistic Relativity versus Natural Partitioning. In S.A. Kuczaj (Ed.), *Language Development: Vol. 2. Language, Thought, and Culture*. pp. 301-304. 1982.
- Gentner, D. Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff, (Eds.) *Action meets word: How children learn verbs* (pp. 544-564). Oxford University Press. 2006.
- Glennen, S. & DeCoste, S. *Handbook of Augmentative and Alternative Communication*. San Diego, CA: Singular Publishing Group, Inc. 1997.
- Gibson J.J. A Theory of Pictorial Perception. *Educational Technology Research and Development*. 2(1). 1954.
- Gibson, J. J.*The Perception of the Visual World*, Boston: Houghton Mifflin. 1950.
- Goodglass, H. *Understanding aphasia*. Academic Press. San Diego, CA, US: 1993.
- Goodglass H and Geschwind N. Language disorders (aphasia). In Carterette EC and Friedmann MP (Eds), *Handbook of Perception*. New York: Academic Press, 1976.
- Google Images. http://images.google.com/. 2008.
- Gillette, J., Gleitman, H., Gleitman, L., and Lederer, A. Human Simulations of Vocabulary Learning. *Cognition*. 73(2) 135-176. 1999.
- Guyot, F. *Etude de la perception sonore en terme de reconnaissance et d'appreciation qualitative: une approche par la categorization*. Ph.D. thesis, Universite du Maine. 1996.
- Hardin Meta Directory of Internet Health Sources, Iowa University.
  http://www.lib.uiowa.edu/hardin/md/. 2009.
- Hartmann, W. M. Sounds, signals, and sensation: Modern acoustics and signal processing. New York.: Springer Verlag. 1997.
- Hazel, G. and Cockerill, H. Communication Resources. *Communicating without Speech: Practical Augmentative & Alternative Communication*. High Holborn, London: Mac Keith Press. 2001.
- HealthCentral.com 3D medical animation library.
  http://www.healthcentral.com/animation/408/index.html. 2009
- Hearst, M. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. *Proceedings of the 7th Annual Conference of the University of Waterloo Center for the New OED and Text Research*. 1991.

- Hill, K. and Romich, B. A Rate Index for Augmentative and Alternative Communication. *International Journal of Speech Technology*. 5(1), 57-64. 2004.
- Hirst, G. and St. Onge, D. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. Fellbaum,C.(ed.), *WordNet: An Electronic Lexical Database*. 1998.
- Hochberg, J. The Psychophysics of Pictorial Perception. *Audio-Visual Communication Review*. 10, pp. 22-54. 1962.
- 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.
- Hough M.S. Categorization in Aphasia: Access and Organization of Goal-derived and Common Categories. *Aphasiology*. 7(4) 335-357. 1993.
- Houix, O., Lemaitre, G., Misdariis, N., and Susini, P. Closing the loop of sound evaluation and design (closed) deliverable 4.1 everyday sound classification part 2 experimental classification of everyday sounds. (4.1 Part 2), June 1 2007.
- Houix, O., Lemaitre, G., Misdariis, N., Susini, P., Franinovic, K., Hug, D., Otten, J., Scott, J., Visell, Y., Devallez, D., Fontana, F., Papetti, S., Polotti, P., and Rocchesso, D. Closing the loop of sound evaluation and design (closed) deliverable 4.1 everyday sound classification: Sound perception, interaction, and synthesis part 1 state of the art. (4.1 Part 1), June 1 2007.
- Houts. P., Doak, C. Doak, L., and Loscalzo, M. The Role of Pictures in Improving Health Communication: A Review of Research on Attention, Comprehension, Recall, and Adherence. *Patient Education and Counseling*. 61(2): 173-190. 2006.
- Howard, J.H. and Ballas, J.A. Syntactic and Semantic Factors in the Classification of Nonspeech Transient Patterns. *Perception and Psychophysics*. 28(5):431-439. 1980.
- Huer, M.B. Examining Perceptions of Graphic Symbols Across Cultures: Preliminary Study of the Impact of Culture/Ethnicity. *Augmentative & Alternative Communication*.16,180-185. 2000.
- Huggins, W.H. and Entwisle, D.R. *Iconic Communication: an Annotated Biography*. The John Hopkins University Press, Baltimore, MD. 1974.
- Huttenlocher, J and Lui, F. The Semantic Organization of Some Simple Nouns and Verbs. *Journal of Verbal Learning and Verbal Behavior*, 18: 141–162, 1979.
- Internet Usage Statistics. http://www.internetworldstats.com/. 2009.
- Jenkins, J. J. Acoustic information for objects, places, and events. *Persistence and change: Proceedings of the first international conference on event perception*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1985.
- Jiang, J. and Conrath, D. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings on International Conference on Research in Computational Linguistics*. 1997.
- Kangas, K. A. and Lloyd, L.L. Augmentative and Alternative Communication*. Human Communication Disorders: An Introduction, (5th ed.)* Shames, G.H., Wiig, E. H., and Secord, W. A. ed. Boston: Allyn and Bacon. 1998.
- Katz, M., Kripalani, S., and Weiss, B. Use of Pictorial Aids in Medication Instructions: A Review of the Literature. American Society of Health-System Pharmacists, 63: 2391-2397. 2006.

- Kauhanen, M.L, Korpelainen, J.T., Hiltunen, P., Määttä, R., Mononen, H., Brusin, E., Sotaniemi, K.A., and Myllylä, VV. Aphasia, depression, and non-verbal cognitive impairment in ischaemic stroke. Cerebrovasc Dis;10:455-461. 2000.
- Keeler, M. The Place of Images in a World of Text. *Computers and the Humanities*. 36 pp. 75-93. 2002.
- Kennedy J.M. A Psychology of Picture Perception: Images and Information. Jossey-Bass Publishers. 1974.
- Kilgarriff, A. BNC database and word frequency list: http://www.kilgarriff.co.uk/bnc-readme.html. 2008.
- Kjelldahl, L. A. Survey of Some Perceptual Features for Computer Graphics and Visualization. *In Proc. Linkoping Electronic*, 2003.
- Kleinman, A., Eisenberg, L., and Good, B. Culture, Illness, and Care: Clinical Lessons from Anthropologic and Cross-Culture Research. *The Journal of Lifelong Learning in Phychiatry*, 4(1): 140-149. 2006.
- Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N. and Neuhoff, J. "Sonification Report: Status of the field and research agenda", Technical Report, ICAD, 1999.
- Kretsedemas, P. Language Barriers and Perceptions of Bias: Ethnic Differences in Immigrant Encounters with the Welfare System. *Journal of Sociology& Social Welfare*. 2005.
- Kunkler-Peck, A.J. and Turvey, M.T. Hearing Shape. *Journal of Experimental Psychology: Human Perception and Performance*. 26(1):279-294. 2000.
- Lahey, M. and Bloom, L. *Language Disorders and Language Development*. Allyn & Bacon, Inc. 1988.
- Lakatos, S., McAdams, S., and Causse, R. The Representation of Auditory Source Characteristics: Simple Geometric Source. *Perception&Psychphysics*. 59(8):1180-1190. 1997.
- Lakoff, R. Women's Language. *Language and Style*. 10(4), 222-247. 1977.
- Landauer, T.K. & Dumais, S.T. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,*104,211-240. 1997.
- Language Master. http://www.exspeechprod.com/LMpage.html. March 2010.
- Lanners, E. *Illusionen.* Lucerne: Verlag C. J. Bucher, 1973.
- Leacock, C. and Chodorow, M. Combining Local Context and WordNet Similarity for Word Sense Identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. 1998.
- Leiner, M., Handal, G., Williams, D. Patient Communication: A Multidisciplinary Approach Using Animated Cartoons. *Health Educ Res*. 19: 591-595. 2004.
- Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., and McAdams, S. The Sound Quality of Car Horns: A Psychoacoustic Study of Timbre. *Acta Acustica United with Acustica*. 95(2):356-372. 2009.
- Lesk, M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of SIGDOC'86*. 1986.
- Ley, P. Satisfaction, Compliance, and Communication. *Br J Clin Psychol*. 21: 241-254. 1982.

- Li, X., Logan, R.J., and Pastore, R.E. Perception of Acoustic Source Characteristics: Walking Sounds. *Journal of the Acoustical Society of America*. 90(6):3036-3049. 1991.
- Lin, D. Using Syntactic Dependency as a Local Context to Resolve Word Sense Ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 64-71. 1997.
- Lingraphica, the aphasia company. http://www.aphasia.com/. March 2010.
- Luce, P.A and Pisoni, D.B. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*. 19(1) 1-36. 1998.
- Lullo, B. AphasiaFox: Implementing Sound and Video. *Independent Work Report (unpublished)*. Princeton University. Fall 2009.
- MacDonald, T. An Augmented Lifestyle. *Augmentative Communication in Practice: An Introduction*, eds. Millar, S. and Wilson, A. CALL Center, Univ. of Edinburgh. 1994.
- Mackay, A.J., Connor, L.T., Albert, M,L., and Obler, L.K. Noun and Verb Retrieval in Healthy Aging. *Journal of the International Neuropsychological Society*. 8(6), 764-770. Cambridge University Press. 2002.
- Makaton Charity. http://www.makaton.org/. 2010.
- Marcell, M.M., Borella, D., Greene, M., Kerr, E., and Rogers, S. Confrontation naming of environmental sounds. Journal of Clincial and Experimental Neuropsychology, 22(6):830–864, 2000.
- Marcell, M.M., Malatanos, M., Leahy, C., and Comeaux, C. Identifying, rating, and remembering environmental sound events. Behaviour Research Methods, 39(3):561–569, 2007.
- Marslen-Wilson, W.D., & Welsh. A. Processing Interactions and Lexical Access during Word -Eecognition in Continuous Speech. *Cognitive Psychology*, I0, 29-63. 1978.
- Matzig, S., Druks, J., Masterson, J., and Vigliocco, G. Noun and Verb Differences in Picture Naming: Past Studies and New Evidence. *Cortex*. 45, 738-758. 2009.
- Mayer, R.E. *Multimedia learning*. Cambridge University Press. 2001.
- Mayer, R. and Moreno, R. A Split-attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory. *Journal of Educational Psychology*, 90(2), 1998.
- Mayer-Johnson. http://www.dynavoxtech.com/. 2009.
- Medindia Animations / Interactives. http://www.medindia.net/animation/index.htm. 2009.
- McClelland, J.L. & Elman, J.F. Interactive processes in speech perception: The TRACE model. In McClelland & Rumelhart (Eds.) *Parallel Distributed Processing*: Volume 2. 1986.
- McClelland, J. L., & Rumelhart, D. E. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407. 1981.
- McFadden, D. AAC in the Community – A Personal Viewpoint. *Widening t he Perspective*, eds. Millar, S. and Wilson, A. CALL Center, Univ. of Edinburgh. 1995.
- MedlinePlus. Speech and Communication Disorders. http://www.nlm.nih.gov/medlineplus/speechandcommunicationdisorders.html. March 2010.
- Microsoft Research Cambridge Image Dataset. http://research.microsoft.com/en-us/projects/objectclassrecognition/. March 2010.

- Mihalcea, R. and Leong, CW. Toward Communicating Simple Sentences Using Pictorial Representations. In *Proc. The 7th Conference of the Association for Machine Translation in the America*. 119-127. 2006.
- Millar, S.V. and Scott, J. What is Augmentative and Alternative communication? In Augmentative Communication in Practice: an Introduction, ed. Allan Wilson; CALL Centre, Univ of Edinburgh, 1998.
- Miller, G.A. and Fellbaum, C. Semantic networks of English. *Cognition*, 41, 1991.
- Minspeak.com. http://www.minspeak.com/. March 2010.
- Mirenda, P. Toward Functional Augmentative and Alternative Communication for Students With Autism: Manual Signs, Graphic Symbols, and Voice Output Communication Aids. *Language, Speech, & Hearing Services in Schools* 34 (3): 203–216. 2003.
- Moore, B. C. J. An introduction to the psychology of hearing. 4th ed. Orlando, FL: Academic Press. 1997.
- Moore, B. C. J. (ed.). Handbook of perception and cognition: Vol. 6. Hearing. 1995.
- Morrow, D., Hier, C., Menard, W., and Leirer, V. Icons Improve Older and Younger Adults' Comprehension of Medication Information. *Journal of Gerontology: Psychological Sciences*. 53B(4): 240-254. 1998.
- Moss, H. and Older, L. *Birkbeck Word Association Norms*. Psychology Press: UK. 1996.
- MRC Psycholinguistic Database, University of Western Australia. http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm. 2006.
- Mrs. Riley, making picture cards. http://mrsriley.com/introduction. March 2010.
- Mynatt, J. Designing with Auditory Icons: How Well do We Identify Auditory Cues? In Proc. *CHI'94*. pp 269-270. 1994.
- Nadel, S. A field experiment in racial psychology. British J. of Psychology, 28, 195-211. 1937.
- National Aphasia Association. http://www.aphasia.org/. March 2010.
- Natural Language Toolkit. http://www.nltk.org/. 2009.
- NIH MedlinePlus. http://www.nlm.nih.gov/medlineplus/. 2009.
- Nikolova, S.S., Boyd-Graber, J., and Fellbaum, C. Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools (in press). *Modelling, Learning and Processing of Text-Technological Data Structures, Springer Studies in Computational Intelligence*. 2009.
- Nikolova, S.S., Boyd-Graber, J.L., Fellbaum, C., and Cook, P.R. Better Vocabularies for Assistive Communication Aids: Connecting Terms Using Semantic Networks and Untrained Annotators. In *Proc. ASSETS09*: 171-178. 2009.
- Nucleus Medical Art: 3D Medical Animations. http://catalog.nucleusinc.com/nucleusindex.php 2009.
- Oldfield, R.C. Things, Words and the Brain. The Quarterly Journal of Experimental Psychology, 1747-0226, 18(4), 340 – 353. 1966.
- Paivio, A., Yuille, J. C., & Madigan, S. A. Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76, 1-25. 1968.
- Panos pictures. http://www.panos.co.uk/. March 2010.

- Patterson, R, and Milroy, R. Auditory warnings on civil aircraft: The learning and retention of warnings. *MRC Applied Psychology Unit*. Cambridge, England. 1980.
- Patwardhan, S., Benerjee, S., and Pedersen, T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. *Proceeding of CICLing2003*, pp. 241-257. 2003.
- Patwardhan, S. and Pedersen, T. Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pp. 1-8. 2006.
- Pedersen, T. and Kolhatkar, V. WordNet::SenseRelate::AllWords - A Broad Coverage Word Sense Tagger that Maximimizes Semantic Relatedness. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations*, pp. 17-20. 2009.
- Pedersen, T., Patwardhan, S., and Michelizzi, J. WorNet::Similarity – Measuring the Relatedness of Concepts. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations*, pp. 38-41. 2004.
- Pettersson, R. *Bilder, barn och massmedia.* Stockholm: Esselte L/iromedel, 1981.
- Pettersson, R. Cultural differences in the perception of image and color in pictures. Educational Technology Research and Development. 30(1), pp. 43-53. 1982.
- Pereverzev, S. V., Loshak, A., Backhaus, S., Davis, J. C., and Packard, R. E. Quantum oscillations between two weakly coupled reservoirs of super fluid, Nature 388, 449-451. 1997.
- pdictionary.com. The internet picture dictionary. http://www.pdictionary.com. 2009.
- Photo.net, a community of photographers. http://photo.net/. March 2010.
- Picasa, free photo editing. http://picasa.google.com/. March 2010.
- Pics4Learning. http://pics4learning.com/. March 2010.
- Pictoworld. http://www.pictoworld.com/. March 2010.
- pidic.com. Picture dictionary. http://www.pidic.com. 2009.
- polarcloud.com. Rikaichan: a popup Japanese-English/German/French/Russian dictionary extension for Firefox. http://www.polarcloud.com/rikaichan/. 2008.
- Prentke Romich Company (PRC). http://www.prentrom.com/. March 2010.
- Proloquo2Go: AAC in Your Pocket. http://www.proloquo2go.com/. March 2010.
- Raaijmakers, J.G.W.; Schiffrin, R.M. Search of associative memory. *Psychological Review* **8** (2): 98–134. 1981.
- Ramsay, C.B., Nicholas, M., Au, R., Obler, L.K., and Albert, M.L. Verb Naming in Normal Aging. *Appl Neuropsychol*. 6(2), 57-67. 1999.
- Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. 1995.
- Richardson, J.T.E. Concreteness and Imageability. *The Quarterly Journal of Experimental Psychology*. 27(2) pp. 235-249. 1975.
- Rogers, Y. Pictorial communication of abstract verbs in relation to human-computerinteraction. *British Journal of Psychology*, 78, 99-112, 1987.
- Rosch, E. Principles of Categorization (Prototype Theory). 1978.

- Rubin, E. *Figure and Ground*. Van Nostrand, New York, NY. 1958.
- Ryan, T, and Schwartz, C. Speed of perception as a function of mode of representation. *American Journal of Psychology*, 69, 193-199. 1956.
- Sandt-Koenderman, M. High-tech AAC and aphasia: Widening Horizons? *Aphasiology*. 18(3), 245-263. 2004.
- Scavone, G., Lakatos, S., Cook, P., and Harbke, C. Perceptual Spaces for Sound Effects Obtained with an Interactive Similarity Rating Program. Intl. Symposium on Musical Acoustics, Perugia, Italy. 2001.
- Schonell, F.J., Meddleton, I.G., and Shaw, B.A. A study of the oral vocabulary of adults. Research Study No. 1, Faculty of Education, University of Queensland Press, Brisbane. 1956.
- Scott, J. Low Tech methods of Augmentative Communication, In Augmentative Communication in Practice: an Introduction, ed. Allan Wilson; CALL Centre, Univ. of Edinburgh, 1998.
- Shane, H.C. Using Visual Scene Displays to Improve Communication and Communication Instruction in Persons With Autism Spectrum Disorders. *Augmentative and Alternative Communication*.15 (1). 2006.
- Silver Lining Multimedia, Visual Essentials. http://www.silverliningmm.com/. March 2010.
- Smith, E.E., Shoben, E.J. & Rips, L.J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 1*, 214-241. 1974.
- Sojournor, R. and Wogalter, M. The Influence of Pictorials on the Comprehension and Recall of Pharmaceutical Safety and Warning Information. *Int J Cogn Ergon*. 2: 93-106. 1998.
- Steele, R.D., Weinrich, M., Wertz, R.T., Carlson, G.S., and Kleczewska, M.K. Computer-based visual communication in aphasia. *Neuropsychologia*, 27(4) 409-26, 1989.
- Storey V.C. Understanding Semantic Relationships. *VLDB Journal*. 2, 455-488. 1993.
- Strauss, J. and McCarthy, J. The Use of Animated Feedback in Visual Scene AAC Display. Poster. Accessed in March 2010.
- Stuart, S. Topic and vocabulary use patterns of elderly men and women of two age cohorts. Doctoral Dissertation, 1991.
- Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. Characterizing the Sound Quality of Air-conditioning Noise. *Applied Acoustics*. 65(8): 763-790. 2004.
- Swaab, T.Y., Baynes, K., and Knight, R.T. Separable Effect of Priming and Imageability on Word Processing: An ERP Study. *Cognitive Brain Research*. 15, pp. 99-103. 2002.
- Symbolstix.com. http://www.symbolstix.com/. March 2010.
- Tactile Symbols. http://www.tsbvi.edu/Education/vmi/tactile_symbols.htm. March 2010.
- Takasaki, T. PictNet: Semantic Infrastructure for Pictogram Communication. *The Third International WordNet Conference* (GWC-06) 279-284. 2006.
- Tangoramas. http://www.spectronicsinoz.com/library/tangoramas. March 2010.
- 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.
- Turnbull, C. *The forest people*. New York: Simon and Schuster (Clarion Book), 1961.

- Tversky, B. and Morrison, J. Animation: Can It Facilitate? *International Journal of Human-Computer Studies*. 57: 247-262. 2002.
- Tzanetakis, G. and Cook, P. Musical Genre Classification of Audio Signals. In Proc. IEEE Transaction of Speech and Audio Processing. 10 (5), 293-302. IEEE Press, 2002.
- University of Pennsylvania Health System. http://pennhealth.com/health_info/animationplayer/. 2009.
- University of Texas Medical Branch Health Animations. http://www.utmbhealthcare.org/. 2009.
- 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.
- Van der Veer, N. J. Confusion Errors in Identification of Environmental Sounds. Paper presented at the Meeting of the Acoustical Society of America, Cambridge, MA. 1979.
- Van der Veer, N. J. Ecological acoustics: Human perception of environmental sounds. Dissertation Abstracts International. 40/09B, 4543. 1979.
- Van Hell, J. and De Groot, A. Conceptual Representation in Bilingual Memory: Effects of Concreteness and Cognate Status in Word Association. *Bilingualism*, 1(3),193-211. 1998.
- Vanrie, J. and Verfaillie, K. Perception of biological motion: A stimulus set of human point-light actions. *Behavior Research Methods, Instruments, and Computers*, 36(4): 625-629, 2004.
- Von Ahn, L. and Dabbish, L. Labeling Images with a Computer Game. In *Proc. CHI 2004*, 319-326. ACM Press, 2004.
- Widgit Symbols Development Project. http://www.widgit.com/widgitrebus/. 2002.
- Wilkinson, K. and Jagaroo, V.Contributions of Principles of Visual Cognitive Science to AAC System Display Design. *Augmentative and Alternative Communication*, 20 (3). 2004.
- Woodbury, K. and Macomber, P. The Theory Behind and Instructions for Creating Visual Scene Display. Tobii Assistive Technology, Inc. & Practical Assistive Technology Solutions. 2008.
- WordNet, an online lexical database. http://wordnetweb.princeton.edu/. March 2010.
- WordNet 3.0 Database Statistics. http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html. March 2010.
- Wren, S. Vocabulary. http://www.balancedreading.com/vocabulary.html. August 2003.
- Wright, A. *Pictures for Language Learning*, Cambridge University Press, U.K., 1990.
- Wu, Z. and Palmer, M. Verb Semantics and Lexical Selection. *Proceedings of ACL*, pp 133-138. 1994.
- Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. 1995.
- Yazdan, M (Edt.). *Multilingual multimedia: bridging the language barrier with intelligent systems*. Intellect Books, 1993.
- Yorkston, K., Honsinger, M., Dowden, P., & Marriner, N. Vocabulary selection: A case report. Augmentative and Alternative Communication, 5, 101-108. 1989.
- Yost, W., Popper, A, and Fay, R. *Auditory Perception of Sound Sources*. Springer. 2007.
- YouTube. http://youtube.com. March 2010.

Appendix A: Sentence List in Section 5.2.3

The underlined words are the verbs tested in the Visual Verb Study II (Cued Comprehension), and words marked with * are the ones represented by a single image (not for testing).

1. Let the lady* get her coat*.
2. She went to the hospital* to see her doctor*.
3. Don't play game* with me.
4. Give me the knife* and you can take the fork*.
5. I am reading your note*.
6. He loves his wife*.
7. I bet he won't ask the question*.
8. I feel happy* about the cellphone*.
9. I can help you to write the book*.
10. She made the strawberry* icecream*.
11. He needs an eye* surgery*.
12. Keep the chocolate* in the fridge*.
13. Shut the window* when you leave the office*.
14. I come to say goodbye*.
15. I want to find my key*.
16. Could you pick a doctor* for me?
17. He put on his hat* and left.
18. He is looking at the picture* of the house*.
19. I remember sitting behind a lady* with white* coat*.
20. I play football*.
21. He helped me with my pants*.
22. I stayed in bed* and could hear the snow* outside the window*.
23. How much did you pay for the car*?
24. Could you bring me a fork* and knife*?
25. I think you should get a computer*.
26. Can I use your cellphone*?
27. You'd better take the pill* to stop the pain*.
28. You should go to bed* now.
29. I try not to eat too much chocolate*.
30. He came back from school*.
31. I don't want you to work at night*.
32. Do you remember where you bought the car*?
33. Don't worry about the computer*.
34. I felt surprised* when the doctor* called.
35. I bet some money* on football*.
36. Do you mind me reading the book* with you?
37. I turned my eye* away from the door*.
38. I didn't hear the cellphone*.
39. Thank you for the icecream*.
40. I saw you picking a white* coat*.
41. Please put away the chair* and shut the door*.
42. I hope that you can stop the game*.
43. Try to sleep* without light*.
44. Could you keep a note* of what he has just said?
45. I stay out in the snow*.
46. Thank you for bringing in my key*.
47. I need a bed* to sleep* on.
48. I start to worry about his eye* surgery*.
49. I've been waiting at the bus stop* for a day*.
50. I am working on a book* about how to use computer*.
51. I am wondering where the hospital* is.
52. I've already paid the money*.
53. Look at the note* on the fridge*.
54. I cannot start the game* now.
55. Let me write this down.
56. I love the pants* you bought.
57. I won't mind if you turn off the light*.
58. He gave some flower* to his wife*.
59. I was wondering when you would call me.
60. What do you think of this hat*?
61. Where can I find a gas station*?
62. I hope that I have more strawberry* to eat.
63. Could you ask the lady* for the door* key*?
64. He made the chair* that you are sitting on.
65. I waited in front of the house*.