Title	Improving Youtube for the vision-impaired : a system for casual crowdsourced annotations
Sub Title	
Author	Westra, Elaine Dora(Okude, Naohito) 奥出, 直人
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2014
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2014年度メディアデザイン学 第342号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002014-0342

Master's Thesis
Academic Year 2014

Improving Youtube for the Vision-Impaired: A System for Casual Crowdsourced Annotations

Graduate School of Media Design,
Keio University

Elaine Dora Westra

A Master's Thesis
submitted to Graduate School of Media Design, Keio University
in partial fulfillment of the requirements for the degree of
MASTER of Media Design

Elaine Dora Westra

Thesis Committee:

Professor Naohito Okude	(Supervisor)
Professor Masa Inakage	(Co-supervisor)
Professor Kouta Minamizawa	(Member)

Abstract of Master's Thesis of Academic Year 2014

Improving Youtube for the Vision-Impaired: A System for Casual Crowdsourced Annotations

Category: Design

Summary

The amount of online user-created content is growing everyday with video content as a dominating medium. For vision-impaired users, video content presents a major accessibility issue due to the dependency on visual cues for meaning. As web video becomes an increasingly important source of culture, it is crucial for all people, regardless of level of vision, to have equal access to this embedded meaning.

This paper proposes a new system that uses crowdsourced base of descriptive annotations to create a more meaningful Youtube experience for vision-impaired users through an unobtrusive tool that provides annotations suitable for the high volume, short format nature of web video.

Popscriptive is a browser extension that augments videos with annotations by matching the Youtube video ID to the descriptive information stored in the Popscriptive database. Annotations are collected into the database through Twitter messages sent by amateur annotators who have little to no experience with video editing software and other complicated annotation mechanisms. This open annotation system matches the open video system of Youtube.

A user test featuring 5 vision-impaired participants with varying levels of vision demonstrated the effectiveness of Popscriptive and the annotation delivery method in providing deeper content engagement for vision-impaired Youtube users.

Keywords:

Vision-Impaired, Online Video, Accessibility, Annotation, Crowdsourcing, Audio Description

Graduate School of Media Design, Keio University

Elaine Dora Westra

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. The Internet and the Vision-Impaired User

Richard McManus, the founder of popular web technology blog, ReadWrite[1], coined the term visual web when he wrote about the growing trend of visually appealing websites and web applications and the rising importance of image and video consumption [30]. In the past year, image and video dependent SNS platforms like Twitter-owned Vine[2] and Instagram[3] grew exponentially with Instagram ranking in as the fastest growing network [28] and Vine as the fastest growing mobile application [41]. Facebook also followed suit with implementations to encourage more visual feeds [35]. This visual trend is far from unexpected as the GUI (graphical user interfaces) has long since replaced purely text-based command line interfaces in popularity among mainstream technology users. But for the visually-impaired user, a more visual web can be an inaccessible web — or at least, a more difficult to accessible web that requires dynamic workarounds and adaptation with every released feature or updated version. Beyond the issue of accessible web user interfaces, there is the issue of the visual content itself and the meaning it presents to the visually-impaired. This paper addresses one of the largest visual content online platforms, Youtube, and proposes a new system for crowdsourced video annotation that allows for casual crowdsourced description of user content by sighted users in order to provide a better content experience for the vision-impaired Youtube user.

In the United States, there are a reported 20.6 million adults suffering from

vision loss, which is about 10% of the national population [48]. Vision loss in this context refers to people who report having blindness or significant difficulty seeing even with a form of corrective lenses [48]. The Internet has functioned as a tool to give these vision-impaired users a way to independently access information that previously had to be acquired person-to-person, resulting in vision-impaired people becoming more technology dependent rather than people dependent – in the same manner as majority of modern society [1]. But as visually impaired people grow more dependent on the Internet, they also become more susceptible to its constant ebbs and flows. With the continuing shift from text-based content to visual-based content, vision-impaired people are faced with the difficulty of finding a new solution beyond existing screen-reading technologies, braille displays, and built-in browser adjustment options.

Currently, these screen reading technologies, such as MacOS's built-in VoiceOver program and the Windows-based JAWS (Job Access With Speech), enable extremely low vision and blind users to hear the content and structure of a webpage read verbally. JAWS remains the most popular screen reader, but its popularity has decreased significantly over the years [49]. Conversely, there has been a steady increase in users of mobile screen readers, with a majority of the screen reading population preferring to use VoiceOver on their iOS mobile device [49]. Android's TalkBack followed at a great distance to rank in with the second most users [49]. Users have expressed a relative amount of confidence towards the recent and ongoing improvements to assistive technology such as screen readers. Instead, they believe the responsibility falls upon site authors to create better, more accessible websites and content [49]. Refreshable braille displays offer a haptic alternative to audio screen readers for vision-impaired users who are literate in braille. Refreshable braille displays commonly feature a row of 18 to 80 cells with 6 to 8 retractable pins that can move up and down to dynamically create braille characters that can be read by a user tracking their fingers over the cells [26]. Other vision-impaired users have a lower level of vision impairment that does not require them to use screen reading technologies. Vision assistance can be satisfactory provided through contrast or sizing adjustments. Because all browsers have built-in functionality to allow for sizing and scaling of page content, it is unnecessary for site authors to replicate this through additional accessibility options [49].

The World Wide Web Consortium[4] is an international organization led by the original creator of the Web, Tim Berners-Lee, concerned with the creation of Web standards. These Web standards are applied by developers, designers, and web content authors throughout the world. Although the official W3C Content Accessibility Guidelines[5] state certain criteria for universal accessibility, most websites fail to meet even the most basic of these guidelines [36]. The W3C Web Content Accessibility Guidelines focus on four main principles: perceivable, operable, understandable, and robust[6]. One of the most basic of these guidelines pertaining to perceivability entails providing text alternatives for non-text content. This is most commonly achieved in the form of alt tags for images and graphs. Alt tags are meant to communicate the content and functionality of an image rather than describe the image itself. In fact, descriptions of decorative images and aesthetics of the site often can become superfluous information to assistive technology users who wish to access the more pertinent content of the site. For video, the guidelines require audio description for relevant visuals to the extent that the visuals are necessary for understanding the content. The W3C is also very strict in not permitting the media content to be made accessible by the community, i.e., through crowd-sourcing volunteer efforts or something similar, as the content should be accessible when published except for in the circumstances of time-based media.

Content on the Internet has a distinguishing feature versus traditional media in the sense that it can be created by anyone to be accessed by anyone. Although the W3C strongly supports this basic sentiment, due to the sheer amount of content that is uploaded every day, it can become unfeasible to expect a large majority of users to follow a strict set of guidelines about content. In the case of video, a quick perusal of Youtube or a tap through Vine will immediately reveal the general lack of visual descriptions provided by the amateur content authors in either the video data or the page content. A study by IBM Research Tokyo showed that for Japanese users, 0% of videos were found to have audio description, compared to .9% for movies and 5.6% for public television [22]. In addition, although web video platforms like YouTube provide some annotation tools and specific features for transcripts and captioning, there are no existing designations for visual descriptions. And with hundreds of hours of video uploaded to YouTube

each minute[7], it is very difficult for vision-impaired users to navigate through this constant flow to find meaningful content with the current lack of assistance from content authors, content publishers, and the web community.

What can be defined as "meaningful content" to a user differs based upon a user's background, preferences, and personality. However, it can be said that there is an overall agreement amongst the online community that meaningful can often being equated with interesting or entertaining enough to be worth viewing or sharing. Popular content or viral videos can be classified as being meaningful content under this definition and be rated based on the number of views or shares through varying SNS channels. The most popular video of all time according to the view count on Youtube is PSY's *Gangnam Style* with almost 2 billion views.[8] This surpasses the second most viewed video, pop idol sensation Justin Bieber's music video for his hit song *Baby*, by nearly a billion views.[9] Although music videos prove to be popular with online audiences based on their audio content alone, in the case of the Korean language *Gangnam Style*, a substantial portion of the international success can be credited to the visual features of the comedic video — most particularly the famous equestrian style dance. Although there is always significant value in audio content, the lack of information from the corresponding visuals can cause a video to lose meaning for the vision-impaired user, as well as discourage the user from pursuing video content that does not already hold some clearly defined audio value based from the user's prior experience.

Vimeo[10], NicoNicoDouga[11], and DailyMotion[12] are a few of the heavyweights in the social web video platform sphere, but no service can compare to the traffic of Google-owned Youtube, the top video-sharing website for user-created content with more than one billion unique views per month.[13] Although Youtube has branched out into the produced content market to compete with services such as Hulu[14], Amazon Instant Video[15], and Netflix[16], the platform is still most well-known for it's user-created content which studies have shown tend to make up around 50% of user-engaged content on the site [7]. User-created content differs from highly produced content in the sense that it is not backed by a major studio or financier and is instead made by a single amateurs or a small group of amateurs. Although there is no widely accepted definition for user-created content, in this paper the researcher will allow the inclusion of all content "which reflects a certain

amount of creative effort." [50] Most user content is created as a hobby or without intentions of making a significant amount of money, however this model has shifted recently with the integration of Youtube's Partner Program[17] which allows users to monetize their content through advertising revenue. However, although there exists monetary incentive in attracting users, because of lack of guidelines or process to encourage content authors to make their content accessible to the impaired, including the vision-impaired, professional content authors are in the similar state of indifference or ignorance as their amateur YouTube peers.

1.2. Proposed Popscriptive System

The researcher proposes a crowdsourced system to provide contextual descriptive annotations for web video content on Youtube. Web content in all mediums should be accessible by any user regardless of their abilities, age, economic situation, education, geographic location, language, etc.[18], but accessibility alone is not necessarily sufficient to ensure the content can be experienced in a meaningful way. Our system will currently focus exclusively on Youtube as it has been and remains to be the largest platform for web video featuring both professionally produced and user-created content. In addition, the researcher has selected to focus on vision-impaired users as Youtube presently only provides captioning support for hearing-impaired users and users who do not understand the original language of the video.[19] Existing annotation features are designed for sighted users and primarily used for promotion purposes[20].

As an in-browser extension, Popscriptive aims to enhance Youtube for vision-impaired as a tool providing crowdsourced brief descriptive annotations for video content. An in-browser extension is an optional enhancive feature that can be downloaded separately through the web stores of a particular browser, in this instance, Google Chrome. An extension can supplement normal browsing behaviour by using Javascript, along with HTML and CSS, to inject new content into the page between existing page content, to modify existing page content, or provide related content in a separate tab or pop-up. The benefit to an extension is that is does not require a user to utilize an entirely new or separate SNS, application, or platform in order to complete their desired goal of content retrieval and con-

sumption. It was a key point of importance in this research to avoid creating a separate web service that would exclusively cater to vision-impaired users, in effect isolating them from the sighted Internet community and possibly causing rejection of the assistive system [44]. The goal of the research was to provide a better experience with an existing popular content platform in order to enable a more open, more accessible, and more meaningful Internet where content can be understood and enjoyed by any user, regardless of their level of vision.

The extension itself will be built with simple Javascript and HTML that refers within a pop-up to a basic web application that can handle more complicated database communication. The web app will also use Mozilla's Popcorn Javascript Library project to inject annotations retrieved from the database into the Popcorn Javascript in order to provide brief descriptive text based on the time stamp of the played video. The video itself will be embedded from Youtube based on the URL of the current Youtube video page when the extension is in an enabled state. The extension itself is a Page Action and will only appear on valid Youtube pages that match a specified URL pattern.

As it was desirable to minimize the separation of the user from the existing platform, Youtube, the extension allows for almost all user activity to be conducted normally through Youtube itself. Search and recommendations, along with comments and user profile data, will all be provided by Youtube. The extension only is enabled when a user is browsing an individual video page and wishes to see a descriptive text in order to attain a deeper comprehension of the video context. After watching the video with descriptive text, the user can then return directly to their Youtube experience. They can read or add comments to the video page, continue on to recommended videos, or begin a new search in a seamless transition that requires only a single exit out of the pop-up that was created by the extension for the particular video viewed.

Annotations themselves are provided without monetary compensation by an online community of casual, non-professional volunteers. As professional annotations are time and money intensive, it would be unfeasible for professional-level annotations to be provided for everyday user content, especially considering the massive amounts of data uploaded to Youtube every hour of every day. In a similar vein, it is also unreasonable to expect content authors to provide descrip-

tive annotations for their own content as there currently exists no tool or basic encouragement to provide such information through the Youtube uploader and channel management. To follow the example of the Youtube community platform itself, the researcher chose to use a non-professional source for annotations through the crowdsourcing of interested parties (for example, friends and family of vision-impaired users or individually motivated casual volunteers). To prevent presenting an additional barrier in the form of a separate web service or platform that might intimidate potential annotators, the researcher decided to use popular social networking service Twitter as the bridge between annotators and the annotation database behind the Popscriptive extension. Albeit a significantly smaller number, Twitter's 255 million active users[21] is still a comparable statistic to Youtube's 1 billion unique user views per month. Twitter-based annotators tweet at the designated Popscriptive Twitter account with a video ID, start and stop timestamps, and a brief description. With the correct tweet formatting, the database will be able to automatically sweep the Twitter account in order to add annotations to the database based on the video ID information. As Twitter limits all tweets to be within 140 characters, this will provide the additional benefit of restraining annotators to writing brief descriptions, as it has been found in other studies that lengthy descriptions can lead to overlapping and distract the viewer rather than enhance the experience [12].

The crowdsourced annotations along with the in-browser extension will create a total system, Popscriptive, that will provide the visually-impaired Youtube user with a more contextual, and therefore enjoyable, video-watching experience. The user will be able to use their screen-reading technology to access the provided brief text descriptions at any frame in the video in order to understand important visual cues in the scene.

The structure of this thesis consists of five main chapters that build from an overview of the situation of assistive technology and the Internet for visually-impaired users and a literature review of existing works and research related to the current effectiveness of Internet accessibility guidelines and interfaces, entertainment accessibility offline and online, and crowdsourcing as a solution. Chapter 3 introduces the conducted fieldwork leading to the formation of the design concept of Popscriptive. A user test and evaluation is discussed in Chapter 4. Chap-

ter 5 concludes the research and provides a prediction of the future direction of accessibility of online video for vision-impaired users and Popscriptive.

Notes

1 http://readwrite.com/

2 http://vine.com/

3 http://instagram.com/

4 http://www.w3.org/Consortium

5 http://www.w3.org/standards/webdesign/accessibility

6 http://www.w3.org/standards/webdesign/accessibility

7 http://www.youtube.com/yt/press/statistics.html

8 https://www.youtube.com/watch?v=9bZkp7q19f0

9 https://www.youtube.com/watch?v=kffacxfA7G4

10 https://vimeo.com/

11 http://www.nicovideo.jp/

12 http://www.dailymotion.com/

13 http://www.youtube.com/yt/press/statistics.html

14 http://www.hulu.com/

15 http://www.amazon.com/Instant-Video/b?node=2858778011

16 https://www.netflix.com

17 https://www.youtube.com/partners

18 http://www.w3.org/WAI/users/Overview.html

19 https://www.google.com/accessibility/products

20 https://www.youtube.com/yt/playbook/annotations.html

21 https://about.twitter.com/company

Chapter 2

Related Works

2.1. Web Accessibility for the Vision-Impaired

The World Wide Web Consortium (W3C) published the first version of their Accessibility Guidelines in 1999. The guidelines were intended for "web content developers [to] follow in order to make pages more accessible for people with disabilities as well as more useful to other users, new page viewing technologies (mobile and voice), and electronic agents such as indexing robots[1]." In addition, the W3C stressed that accessibility does not equate to minimal UI design, but instead it means "thoughtful" UI design. The guidelines provided are meant to "outline procedures for authors, particularly those using multimedia content, to ensure that the content and functions provided by those elements are available to all users. In general, authors should not be discouraged from using multimedia, but rather should use it in a manner which ensures that the material they publish is accessible to the widest possible audience[2]."

Although the initial W3C Accessibility Guidelines 1.0 may have achieved their goal of increasing awareness about accessibility issues [21], the actual "impact of WCAG 1.0 on improving the accessibility of the Web remained quite low throughout the period of its use," based on both user and automated evaluations [36]. This was most likely due to the fact that web content authors display a relatively low level of knowledge about accessibility tools and guidelines as demonstrated in a study by Lazar et al. that found 22% of site owners had absolutely no knowledge pertaining to accessibility guidelines [24].

In 2008, the W3C revised their guidelines to create Accessibility Guidelines 2.0. However, a comprehensive study involving 30 million web pages conducted in 2010 by Lopes et al. showed that under 4% of elements met the 2.0 Success Criteria [27]. This number is already considerably limited due to the fact that only a proportion of W3C Accessibility Guidelines can be tested through automation. A study by Power et al. that was featured at CHI 2012 (ACM SIGCHI Conference on Human Factors in Computing Systems) investigated the relationship between W3C Accessibility Guidelines and actual user experience [36]. The research included a user study of 32 vision-impaired participants who were asked to carry out various tasks on selected websites and rate errors encountered on a four point scale (Cosmetic, Minor, Major, Catastrophic) [36]. The results revealed that web content authors were not implementing the current version of the W3C Accessibility Guidelines. In addition to the lack of implementation, a deeper problem was found when sites with guideline implementation still failed to indicate that people with vision-impairment would experience fewer problems [36].

The second largest number of problems found in the study revolved around multimedia with audio description. In order to adhere to the W3C Accessibility Guidelines, it is necessary for a site to provide an additional track of audio description or a time-indexed text description alternative. Of the 31 non-enhanced multimedia problems found by users, 51.6% of the websites surveyed were able to pass the basic level of W3C Accessibility Guidelines by providing a text description of video while the other 48.4% completely lacked audio description or any other alternative [36].

Overall, the study showed that the 2008 update to the W3C Accessibility Guidelines was not having the intended positive effect of greater accessibility for vision-impaired users. There was not a significant decrease in problems between basic guideline conformant and non-conformant websites. There was a positive correlation between the number of W3C Accessibility Guidelines Success Criteria violated and the number of problems encountered by the test users, which the study evaluated as indication that "the current WCAG 2.0 priority levels are too crude of an accessibility measure. [36]" 49.6% of problems encountered were actually "addressed by directly relevant" Success Criteria and only 16.7% directly relevant Success Criteria are being actively implemented on websites which

demonstrates that web content authors face difficulties when creating accessible websites [36].

The authors of the study rejected the idea that Accessibility Guidelines do not necessarily need to cover usability. Instead the authors insisted on the need for accessibility and usability to be addressed together as the two are interdependent due to the role of usability in providing motivation for accessibility. Simply stated, there's no reason to access something that cannot be used. The authors continued their argument by using the results of their study to stress the importance of shifting from a problem-based paradigm to a design-principle-based paradigm that focuses on the people themselves, rather than just on the problems they encounter.

A similar user study was carried out by Ferreira et al. and the results presented at *The 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion* (DSAI) in 2012 [15]. With the motivation that all user interfaces should be universally accessible to every person, independent of their physical, perceptual-motor, social, and cultural background, the authors of the study explored the development of websites that are easily understandable and navigable for visually-impaired users. In a similar manner to the conclusions of Power et al., the paper states that usability issues, rather than simple code compliance, must be emphasized when designing accessible web applications. Although interfaces often rely on a visual presentation, to be usable and accessible for the visually-impaired users the interface must ensure "transparent communication" where the interactions remain user-friendly through the assistive technology so the user is only required to focus on the actual task. Usability issues tend to occur because (1) accessibility is emphasized rather than usability, (2) testing often is done through automated programs resulting in limited detection of issues, (3) the users mental models are overlooked, as users are active — not passive — beings and will use logic to interact with websites [15]. NFR (Non-Functional Requirements) Usability must be incorporated into the system's definition, requiring knowledge by the system engineer. The paper shows that it is necessary to consider usability issues rather than only code compliance with existing accessibility guidelines.

In the paper "User-Sensitive Inclusive Design" by Newell et al., the authors

focus on the need to develop empathy for disabled users in favor of "inclusivity" rather than "universability" as a more achievable goal. And, likewise, "user-sensitive" rather than "user-centred" because "it is rarely possible to design a product that is truly accessible by all potential users [32]." The paper raises the statistic from a previous study by Hocking which reported that 56% of all assistive technology is quickly abandoned and 15% is never used [19]. One reason for this, the authors argue, is the overemphasis on functionality which drives developers to focus on the product rather than the users. Another reason may be the misconception that incorporating consideration for disabled users equals "abandonment of novel and beautiful concepts" [42]. The authors push for appealing design for assistive technology, rather than invisible design – in a manner similar to how eyeglasses have moved from assistive to fashionable [32]. In order to develop empathy for the user, the study proposes using less traditional methods of user studies, such as ethnography and techniques borrowed from professional theatre for a performance of the user, versus the "two-way mirror" style of classic laboratory usability experiments.

As an alternative to existing W3C standard guidelines, many researchers have focused on designing new structures for web content delivery. Pauli Lai proposed a new system for dividing websites into number-assigned logical sections under descriptive headings in order to create an IVRS (Interactive Voice Response System) structure enabling vision-impaired users to access websites by mobile phone in a similar manner to existing automated telephone systems [23]. The proposal emphasized the importance of "semantic elements" with relationships, such as descriptive relationships, represented through a DOM-like model.

Similarly, Rajapakse et al. proposed a set of guidelines for a AUI (Audio User Interface) to provide direct auditory output of content to the visually-impaired user, rather than the existing GUI interpretation through assistive technology [38]. The research emphasizes the importance of the 2D audio environment with spacial positioning for vision-impaired through a comprehensive usability study to support the proposed guidelines and show new design aspects. The current assistive technology of screen readers allow vision-impaired users to access text content in a sequential manner but lack a method for providing complex structures and important spatial information. Previous research involved a similar use of

spatialized sound to enhance usability of 3D modeling applications, but the same concept was found to be viable for 2D environments as well [31]. The advantage of spatialized sound is it allows the user to understand complex space relations between page elements and avoid an overload of the memory caused by top-to-bottom reading.

The design of the prototype took interviews and questionnaires with visually-impaired users into consideration, along with guidelines provided by the Web Content Accessibility Guidelines (1.0). The prototype was tested with vision-impaired users in order to see the effectiveness of a direct AUI with spatial information versus the traditional GUI translated by a screen reader. Although the results of the user study showed the linear solution allowed for a more effective use of time, less errors, and less assistance required, the results reflected users' prior experience and familiarity with using the linear solution. A number of guidelines were devised based off observations that differ from traditional GUI-based guidelines in order to create a two dimensional audio environment, including the importance of a minimal number of interfaces and key strokes to complete a task [38].

Another AUI usability study was conducted by Ashok et al. proposed a Speech-Driven Web Browsing system in order to create an alternative method for browsing that did not require fatigue-inducing keyboard shortcuts and clicks [2]. The findings of the user test with 24 vision-impaired participants resulted in a baseline for evaluating usability of speech interfaces. Users were asked to complete standard web tasks using spoken commands of their choice. Using Wizard-of-Oz experiment tactics, the user's commands were executed by the research team to create the effect of a fully developed and implemented software system and user interface. Free high-level spoken commands allowed users to avoid "mundane and tedious low-level operations" including clicking, tabbing, and typing [2]. The results of the usability study produced an analysis between three different user interfaces (free speech, keyboard, and a combination of the two), validation and desirability of the proposed system. The findings allowed the authors of the paper to create a baseline for speech-enabled web browsing that can be applied to speech-based user interfaces.

Dinesh et al. analyzed accessibility in the context of rural India where illiteracy presents a problem for Internet usage, proposing the concept of re-narration [10].

Although the research focused on an interface system for the "print-impaired" rather than the vision-impaired, there was a marked similarity in the inclusion approach to making content accessible for those who cannot read or understand language presented on a screen. The system used a combination of a web framework, filters, and server-supported browser extensions for a structure based on the semantic social web model. However, rather than rely exclusively on the tool, the "the re-narration activity subsumes the tool aided activity by including a group of narrators who are interested in the community [10]." Re-narration uses crowdsourced information from users to adapt "access to web-content in ways that are relevant for any user, but may be particularly useful to print-impaired users and others who are lost in translation [10]."

Rather than focus on overall structure for web accessibility, Popscriptive aims to focus on making a tool that would include a group of interested narrators in a similar manner to the Re-Narration project. Crowdsourcing will be discussed in further detail in section 2.4. In addition, Popscriptive concentrates on specific content and its meaning, rather than on the structure, interface, or code compliance of the pages holding the content. As the study by Power et al. concluded, guidelines for content authors are not enough to guarantee accessibility or usability for users [36]. Although the authors saw the need for shift from the problem-based paradigm to a design-principle-based paradigm, Popscriptive follows the problem-based-paradigm in offering an immediate solution for an existing service that cannot be immediately re-designed from its basic principles. Ultimately, the Popscriptive project hopes to push Youtube to incorporate vision-impaired accessibility into its basic principles of design as one of the main goals of the study. Albeit being a product of the problem-based-paradigm, Popscriptive also follows the user-sensitive inclusive design techniques of Newell et al. [32] to empathically consider the vision-impaired users from the very beginning stages of idea conception, while avoiding the misconception that attractiveness must be sacrificed for accessibility. The annotation qualities of Popscriptive will provide attractive value for both vision-impaired and non-impaired users, creating more searchable and semantic content.

2.2. Entertainment Accessibility for the Vision-Impaired

Zillmann and Vorderer discuss the "unimaginable wealth of entertainment choices" since the beginning of the information age in the 1980s [53]. Entertainment "obtrusively dominate media content" and their reign will continue in the future [53]. The preeminence of media along with an ever-growing public demand make the present time into the "age of entertainment" as never has "so much entertainment been so readily accessible, to so many, for so much of their leisure time as is now, primarily because of the media communications [53]." When Zillmann and Vorderer refer to "accessibility", they are using the word in a general sense rather than in relation to the disabled and impaired user population. As discussed in the previous section 2.1, technology accessibility for visually-impaired remains a growing issue. It is imperative for visually-impaired users to have equal access to technology for the educational, work, and entertainment purposes. Although attention has been focused on accessibility regarding education and work, for example, the DCMP[3] (Described and Captioned Media Program) funded by the U.S. Department of Education provides an online and offline library of over 4,000 free-loan titles, entertainment has a smaller corpus of dedicated research and advocacy despite its importance for human happiness. [18]

Udo et al. focused on the entertainment event of a live fashion show with simultaneous web broadcasting for their research in accessibility for the vision-impaired [47]. The authors were concerned about the accessibility to cultural events and activities for the vision-impaired. The study identified two types of audio description: sporting and accessibility [47]. Audio description initially developed in the context of sports without a conscious effort to provide content for the vision-impaired. Sports commentary was created to be broadcasted through the radio in order to reach an expanded audience of people, but even after the decline of radio and the rise of television, sports commentary continues to provide entertainment value through a combination of "play-by-play, a description of the ongoing action, and colour commentary, a narrative composed of background information and interpretation of action [17]." Colour commentary differs from traditional audio description as it replaces objectivity with emotion. In addition,

audio description of sports exists primarily for fans or those with prior knowledge.

Udo et al. found that audio description outside of sports fell into the category of accessibility, providing "individuals with vision-impairments access to verbal descriptions of some (but not all) visual stimuli [47]." Audio description in this context were found between dialogue of film, television, and live events [47] and were crucial for preventing "social disadvantage, as [visually-impaired participants lacking audio description access] are unable to fully participate in a culture that is heavily saturated by and based on the enjoyment of audiovisual entertainment experiences [33]."

According to the camera lens approach of G. Frazier, the original developer of audio description in the 1970s, and, later, Pfanstiehl and Pfanstiel [43], an approach also used in the guidelines provided by the Audio Description Associates and the National Center for Accessible Media, an audio description describer should be an "objective interpreter and translator of important visual events, costumes, scenes, and effects that cannot be disambiguated through sound. Describers are encouraged to use precise, but highly, descriptive language, a strategy recommended regardless of venue or genre [47]." However, there is a high time and resource cost for high quality audio description. In addition, the production of a traditional third person audio description involves a separate script that translates important visuals, a narrator, and audio recording and editing [14]. For a live show, the describer must be very familiar with the details of an event or show. Some non-profit organizations around accessibility, such as VocalEyes, provide volunteer describers for art galleries, architecture, and theatre productions[4]. Another accessibility method is the "open description" approach where the audio description is incorporated to the actual script of the production [47]. Other theatres offer live or recorded notes and program information or touch/sensory tours before the show where vision-impaired theatre goers can experience the sets, props, and costumes through touch.

A previous study by Schmeidler and Kirchner centered on television has shown that visually-impaired viewers gain and retain more information with audio description [40]. Another study by Fels et al. evaluated the perception of a first-person audio description style versus the traditional third-person style and found that audiences found the first-person style less trustworthy but entertaining [13].

In an additional study by the same authors, the authoritative nature of the third-person narrator is also discussed as limiting interpretation by the audience.

The user study was carried out during one day of a university fashion show where an experienced and "passionate" fashion and theatre student acted as the describer in order to "to provide a description with a focus on entertainment rather than information [47]." Using both a prewritten skit and improvisation, the describer was able to provide narration for around 60% of the outfits shown in the show, due to the fast nature of the model walk and the abundance of outfits. Vision-impaired participants both live and through web streaming gave positive evaluations for the audio description, suggesting that they would be "comfortable conversing with sighted people about the show and that the description provided would facilitate their discussion [47]." There was mixed response about the description style with some participants appreciating the emotion and others preferring more objectivity. However, many participants answered positively about the additional descriptions that were inserted by the describer as personal commentary. Although fatigue of the describer herself was not an issue during the study, it is an important concern for ongoing performances.

Udo et al. conducted another study on theatre productions with an untraditional style of audio description in an attempt to more correctly convey the directors vision rather than simply describing the set, actors, and lighting [46]. Although informal audio description has existed for many years, with friends and family members serving as describers, formal audio description is often required to be "as objective as possible, void of emotionally subjective interpretation, and should describe "relevant visual action imparted by an actors body language, gesture, scene changes, facial expressions, costumes, and other visual aspects and be inserted within natural pauses in dialogue [46]." Response from the user study where 22 visually-impaired people heard a specialized audio description track during a production of Shakespeare's Hamlet were generally positive. Participants enjoyed the congruent style of audio description with the subject matter but negative responses called for more descriptions, such as those "describing entrances and/or exits, desire for more expressions and gestures, and better descriptions to understand location and time of day [46]." A problem experienced by the researchers was limited description time due to the fast pace of the production.

Because the audio description is broadcasted privately to the earphones of individual vision-impaired theatre goers, it is impossible for the actors to adjust their pace to avoid overlapping a description. Even if a slower pace could be achieved, unnatural pauses could affect the experience of other theatregoers. In conclusion, the study found that vision-impaired audiences desire to be entertained when they attend entertainment events such as live theatre and the audio description at entertainment events should therefore be "consistent with that goal" [46]. It may be possible to achieve this through audio description that "fits linguistically, emotionally, and stylistically, with the performance [46]."

The North American box office is more than 10 times the box office of theatre productions, bringing in $10.8 billion USD in 2012[56]. In addition, movie theatre attendance is higher than all theme parks and major U.S. sporting events combined[7]. Approximately 26% of the adult population in the United States (56 million people) visit a movie theatre once a month. Of the 30,000 movie screens in the United States, approximately 18,000 are enabled for captioning and description technologies for impaired viewers[8]. However, another statistic states that only "approximately 200 movie theatres nationwide [offer] audio description [...] available for first-run film screenings" [43]. In movie theatres, audio description can be provided through the following systems: MoPix[9], DTS Access[10], Fidelio[11], Sony Digital Camera Entertainment Access[12]. With the rise of digital cinema, audio description can now be delivered in a Digital Cinema Package[13] (DCP) with the movie, soundtracks, and accessibility contents.

Currently all major studios in the United States offer audio description for widely released feature films, although often audio description is not available on DVD releases. Sony, Disney, and Universal have offered audio description on most DVD releases since 2010[14]. The ACB (American Council of the Blind) initiative, The Audio Description Project[15], maintains a list of all English-language films with audio description available. The NTN[16] (Narrative Television Network) is an online source for audio described movies, television, and documentaries using Youtube as a host for the augmented content. NTN receives funding from the U.S. Department of Education and, therefore, provided content skews heavily towards education rather than entertainment. An online organization, The Metropolitan Washington Ear, provides free audio description services for the

"blind, visually impaired, and physically disabled people who cannot effectively read print[17]". Interestingly, another service provided by the organization is dial-in audio of newspaper and magazine of major American publications. The Accessible Netflix Project[18] is an online grassroots movement using the channels of a blog and Facebook group[19] to promote accessiblity for online streaming services such as the organization's namesake, Netflix. The group strives for recognition of the need for accessible interfaces and audio described content for vision-impaired users in order to ensure equal access to media entertainment. Some television channels provide 24-hour described programming, including a variety of entertainment options along with news, documentaries, and other original shows. In Canada, AMI-tv is one such channel that is required to be included by all major television service distributors in their basic package offering.

The introduction of smartphones has also opened up the realm for portable audio description that would free visually-impaired users from depending on theatres. Parlamo[20], an app that was originally created to provide language translations for movies expanded its offering to include audio description. A similar is app is MovieReading[21] which will also support automatic syncing to the movies audio for audio description.

2.2.1 Web Video, audio Description, and Annotations

As Brian Charlson, Chairman of the Information Access Committee, stated in his testimony on behalf of the American Council of the Blind in a 2013 U.S. Senate Hearing on ADA and Entertainment Technologies, "Today you can go to a movie theater or watch television shows with video description. Unfortunately, when you visit Web sites that provide this content, most all of the programming is not accompanied by description because there is no requirement to do so.[22]" Youtube is a web platform for online video content with the highest amount of traffic and the largest amount of content uploads compared to any other video content site[23]. 72% of all videos watched online are watched on Youtube[24]. A study by Burgess and Green found that approximately 50% of user-engaged Youtube content can be classified as "user-created" content [7]. The study simplified the definition of "user-created content to be amateur content or "bedroom, boardroom, or backyard productions" versus the traditional media of television, cinema and music

videos, although the authors were aware of the complexity and convergence of media within the site [7]. A survey of Youtube users by Rotman and Preece also indicated that users identified Youtube as a community of communication and interaction, rather than as a broadcasting platform [39]. Although Youtube's Partner Program[25] enables users to profit from their created content, there is still a large gap between the budget of a Youtube video and the $100 million plus budget of a Hollywood summer blockbuster.[26] A Hollywood film has the potential budget to be made accessible through audio description, at least in its initial theatre run, if not its DVD release.[27] However, there exists little to no motivation, financially or legally, for Youtube user content authors to consider accessibility to their videos. Although Youtube has implemented a system for adding closed captioning and language translations, there is no tool, process, or encouragement for adding accessibility for the vision-impaired.[28] Annotations exist in the form of "easter eggs", interactivity, or pop-up promotion tools[29], but require visual recognition for accessibility – therefore greatly reducing their meaning for vision-impaired users. Current recommendations and initiatives relating to video enrichment include the Web Accessibility Initiatives (WAI)[30] support of different versions of temporal content, Xiph.orgs Ogg[31] open video format, and the HTML Accessibility Task Force's advocation of the HTML5 Media and Track elements to hold several tracks for video[32] but they do not apply to Youtube or accessibility for the vision-impaired directly.

Offline video editing software such as Swift ADePT[33] and Magpie[34] (Media Access Generator) allow for the addition of audio description tracks to video. In addition, research by Gagnon et al. has produced a computer-assisted video description production system, VDManager, that uses computer vision technology to automatically detect elements such as indoors vs. outdoors lighting and actors' faces to augment the video-description process for more efficiency [16]. VDManager also uses text-to-speech technology to create synthesized audio descriptions. Another offline application, LiveDescribe, describes itself as being a tool for amateur audio descriptions creation for video[35]. In fact, a study conducted by Branje and Fels [6] using LiveDescribe demonstrated that amateur produced audio description is a feasible method for rapidly expanding accessibility for vision-impaired users to video. Although the LiveDescribe website shows

an attempt to become an independent online community resource, a browse of the available audio described videos shows that only 17 videos are available and there is no apparent recent activity.[36] A similar software was CapScribe[37], which is no longer available due to its dependency on the deprecated Quicktime 7 API, offered "DIY" description editing for Quicktime and Youtube videos with text-to-speech synthesized audio.

Descriptive Video Exchange (DVX)[38] is another software currently under development that will allow users to audio describe any DVD and share the audio and synchronization data through an online platform based on a client/server model. DVX also aims to create a "wiki-style crowd-sourcing of video description in a completely new way, opening the door to amateur description provided for any video content, and distributed to anyone, anywhere[39]." The project endeavours to evaluate the effectiveness of audio description by amateurs through crowd-sourcing from social networking and online communities and in the future will broaden to support online media including YouTube, iTunesU, and other streamed video by utilizing the available public APIs.

YouDescribe[40] is another audio description tool, developed by the Video Description Research and Development Center (VDRDC). YouDescribe focuses exclusively on making YouTube more accessible for the vision-impaired and offers a completely online platform. To add audio description to a specific video, sighted users may log into the web platform and search for the desired video through the sites YouTube API search. After selecting the video, the user pauses it at the appropriate times to record audio descriptions using the site platform. The audio file is then stored in the server where it can be retrieved by vision-impaired users who search for the same video through the YouDescribe site. On a related note, the creator of DVX and a developer of YouDescribe, Josh Miele, is also an active supporter of accessibility audio description for the vision-impaired through his Twitter account[41]. Twitter advocacy will be discussed further in Chapter 5.1.1: Volunteer Communities.

Research by IBM Research advocated using text-to-speech assistive technology in combination with crowdsourcing to provide greater accessibility to online video content [22]. An in-depth study in Japan showed that amateurs can successfully describe videos. In addition, responses from online survey conducted in the United

States and a face-to-face survey conducted in Japan showed that synthesized audio was considered "Comfortable" or "Acceptable" by most participants [22]. With this data, IBM Research proposed a platform that uses text-to-speech synthesized audio either pre-recorded on a server, server-side synthesized, or client-side synthesized for a lower cost solution to audio description availability. Their proof-of-concept platform consists of an authoring tool, an HTML-based player, and a script repository [22].

A study by Encelle et al. focuses on annotation-based video enrichment with the use of earcons and speech synthesis [12]. The paper proposes a fusion system of audio enrichments to create better accessibility to online video. The audio enrichments consist both of standard synthesized speech and the novel concept of earcons, which are nonverbal audio messages that hold some assigned meaning. Results showed that earcons can be used together with traditional speech synthesis but should be accompanied with explanations.

The enrichment process requires two different users groups: the users who enrich the videos and the users who consume the enriched content. Unlike existing formats and tools that utilize "direct" enrichment, the proposed "indirect" system separates the content and rendering into a two step process which provides room for innovation and collaboration. With this process a video can be enriched with three main elements: visual enrichment (captions, still images, video fragments, etc), audio enrichment (voices, sounds), tactile enrichment (vibration, Braille text) [12]. The system consists of a voting-based priority queue which can be provided to friendsourced enrichment producers, who themselves can invite other collaborators. After enrichment is added, the producer can specify the presentation model and share the annotations. In addition, visually-impaired users can search the system database to find videos that have been previously enriched based on their interests and specific impairments, as well as customize their presentation model to create the most comfortable viewing environment. There is also a feedback system to allow the user to report any issues about the enrichment to producers.

The research investigates the potential of audio notifications, or earcons, to convey information related to videos to move beyond the restrictions of traditional audio description techniques for bi-modal enrichment of video with unity to trans-

fer spatial and temporal information. The evaluation of the system was conducted through a user study. Participants were able to attain a high level of understanding about the video content and also responded that annotations needed to be brief and not overlap with the original soundtrack. Because of the brevity of earcons, participants appreciated the mixed usage and found their meanings easy to learn with an explanation [12].

Popscriptive aims to enhance the existing community and content of Youtube through the addition of annotations for the visually-impaired user. These annotations will be translated to audio or braille, based on the preference of the user. In the interest of keeping pace with the exponentially growing amount of user-created content uploaded each day and to avoid any need for financial expenditure, Popscriptive will rely on user-created annotations to match crowdsourced structure of Youtube. Although video editing tools both online and offline, like Swift ADePT, Magpie, LiveDescribe and DVX, provide a method for adding supplementary audio tracks, in order to avoid imposing any knowledge or time-related hurdles for the user, Popscriptive offers a system that requires no editing task. Instead, the user will only need to type a brief message in a process identical to sending a standard tweet on Twitter. In addition, the user will not be required to use their own voice or go through the process of audio recording – a key part of the process for web tool, YouDescribe. Text-to-speech synthesized audio or text-to-braille has the potential to provide description access for a greater amount of content as discussed in the individual studies of Kobayashi [22] and Branje and Fel [6]. One of the Popscriptive prototypes also incorporates the use of earcons in a similar manner to Encelle et al. [12] in order to indicate the availability of a description for a Popscriptive during the video.

2.3. Crowdsourcing

As mentioned in the previous section, crowdsourcing was found to be an effective method for providing annotation for an expansive amount of online video content by Kobayashi [22] and is currently being incorporated into the models of DVX, LiveDescribe, and YouDescribe. Crowdsourcing involves using a number of people, paid or unpaid, to complete a given task or solve a certain problem.

The concept of crowdsourcing can be described in the terms of "peer production, user-powered systems, user-generated content, collaborative systems, community systems, social systems, social search, social media, collective intelligence, wikinomics, crowd wisdom, smart mobs, mass collaboration, and human computation" and can occur both in the physical and the digital world [11]. The main challenges faced by crowdsourcing systems consist of (1) recruiting contributors, (2) contributors level of skill, (3) combination of contributed material, (4) managing abuse, and (5) balancing openness versus quality [11]. One of the key applications of crowdsourcing is for rapidly building databases [11]. Author of *Crowdsourcing*, Daren C. Brabham, defines crowdsourcing as a deliberate blend of bottom-up, open, creative process with top-down organization goals [5].

Vondrick et al. conducted a three year study by on using crowdsourcing to efficiently and economically annotate video for pure monetization purposes, not related to accessibility for vision-impaired users [51]. The research involved creating the user interface for open platform VATIC[42] (Video Annotation Tool from Irvine, California) to allow annotators to label video content at optimum levels of quality and speed. The study results showed that macro-tasks using specialized workers rather than traditional micro-tasks using generic crowdsourcing was more efficient for complex video annotation. The goal of the research was to inspire a greater interest in developing massive labeled video data sets for data-driven computer vision applications as image labeling has previously been proven to be crowdsourcable for the same purpose. However, although a similar amount of video data exists, the labeling process has not been as successful due to the dynamic nature of changing frames. Although some tools, such as FlowBoost[43], allow for video annotation to build large data sets, they do not emphasize an economical workflow. The paper proposed a new platform for large scale, high quality, and economical video annotation. Based on user studies, the user interface was designed for efficiency with constrained choices and more simplicity. Annotators were crowdsourced from Amazon Mechanical Turk[44], an online marketplace that connects available workers with employers with for on-demand remote completion of small tasks, using a "golden standard" method to filter out lower producing workers by presenting a difficult pre-task [51].

The designed system uses a fixed key frame schedule versus user-defined keyframes.

Although user-define keyframes allow for more precision and flexibility, it comes with the consequence of greater time expense. The fixed key-frame system pauses the video at a given interval to request annotation updates from the annotator. Participants in the study found the automatic pause schedule to be helpful in the work process. It is essential to set an appropriate interval frequency for high quality labels. Although user-defined keyframes can help to set an appropriate frequency, user study results showed that efficiency was 33% lower under the user-set workflow [51]. The studies also found the single-object approach, rather than a all-object or group-object approach, to be the most effective. Although seemingly counterintuitive, users found the single object process to be more efficient and preferable. The limited features of the interface (fixed key frames, predefined objects, support for only rectangular bounding boxes) is key to its success because the worker has a smaller number of choices and can be contained in a closed world environment for lower anxiety and higher efficiency. Because initially users attempt to accomplish too many tasks simultaneously, the constrained interface lends itself to a one decision at a time process for more efficient video annotation.

In conclusion, the research showed that it is necessary to build intelligent annotation protocols using tracking and interpolation in order to attain high quality and economical labels for a platform that can annotative massive data sets. Crowdsourcing, especially if contributors are not annotation experts, cannot be relied on alone for progress in video data set annotations for improved computer vision without intelligent annotation protocols.

Research by Sanjana Prasain utilized crowdsourcing to provide information to vision-impaired bus riders for assistance in finding their appropriate stop based off of surrounding landmarks [37]. Visually-impaired people cannot drive, so many depend on public transportation systems. Therefore, the usability of public transportation systems is extremely important. Previous work on the GoBraille [3] project showed that finding exact locations of bus stops was a struggle for vision-impaired commuters. In addition, the vision-impaired commuters stated that they preferred to avoid using specialized devices like Braille notetakers. Prasain took both these factors into account to create a system to improve the usability of the public bus system.

StopFinder is a system to provide information about landmarks around bus

stops in order to assist vision-impaired commuters in finding the stop [37] . The system utilizes crowdsourcing to provide the localized information from a "crowd" network of other public transportation users who provide information to the database. As vision-impaired commuters are unable to read signs to attain information about bus stop locations, the crowdsourced information provides alternative information such as street names, direction to walk from the intersection, and the objects in the immediate surrounding vicinity of the bus stop.

Although there are applications available to improve the public transportation usability for commuters, many of these applications are not made accessible for vision-impaired users. Consequently, many of these applications also fail to address problems that vision-impaired commuters face. StopFinder focused on providing information needed by vision-impaired commuters through an interface designed with vision-impaired users in mind. Various landmarks are situated in the vicinity around bus stops, including shelters, bunches, garbage cans, street texture, grass, poles, coffee shops, restrooms, and other remarkable things. It was crucial in the research to recognize what landmarks are the most meaningful out of the many landmarks in order to provide the user with the most effective information without overloading them with an excess of landmark information. The application requests the stop information by stop ID to match the entry in the database to return two entries: highest rated and most recently added.

To evaluate StopFinder, a user study was conducted in a lab environment. A user was given a brief explanation about the application, asked to complete a set of tasks, then underwent a semi-structured follow-up interview. The interview was constructed to evaluate the usability of the app, in addition to the sense of independence and safety the user was given in respect to using public transportation with the app. Participants reported high satisfaction with the information provided, as well as an enhanced sense of independence and safety. As information is provided by crowdsourcing from other users, the reliability of the information was a crucial issue, but the participants felt reassured by the rating system as to the reliability of provided information.

Bigham et al. also used crowdsourcing for real-time data in their research involving an application, VizWiz, that enables vision-impaired users to take a photo and ask questions about the content of the photo to remote sighted people on the

web [4]. In order to address an immediate lack of access to visual information, the text-to-speech application functions with a real-time response rate through a layer of abstraction built on top of the Amazon Mechanical Turk API called quikTurkit. quikTurkit reduces the latency by creating a pool of crowdsourced workers ready to offer answers to questions, but there is also a relative financial cost of keeping an available pool. Software using computer vision technology, such as optical character recognition to identify visual features to vision-impaired users, already exists but these products often have few functions, many errors, and a high monetary cost [4]. Instead, humans in the form of friends, volunteers, and workers often provide the solution for visual assistance. The app in combination with the quikTurkit approach provides a human solution augmented by web. One of the future directions for VizWiz is to expand the crowdsource base from Mechanical Turk to include an individual's social network. However, there may be a preference among users for anonymity versus cost.

A later study by the same researchers employed the VizWiz application to enable vision-impaired users to get fashion advice from pre-screened volunteer workers [8]. Due to the subjective manner of fashion, computer vision was deemed unfeasible for providing advice. In addition, the answers provided through an open marketplace crowdsource base like Mechanical Turk could not be guaranteed to be culturally appropriate, sensitive to the user, and private. With volunteer contributors, real-time or timely answers could not be guaranteed due to the lack of organization in coordinated scheduling to ensure at least one volunteer was available at all times of the day. During the user study, participants stated they were more willing to trust the opinion of the screened volunteer than an unscreened volunteer. In general, the experiment results showed that users were comfortable in asking questions to strangers. Trust was built through the volunteers detailed responses and through validated answers.

Popscriptive will use crowdsourcing to provide brief simplified annotations in a casual manner that replicates the user-created feeling of Youtube content, rather than at the professional frame-by-frame level demanded by Vondrick et al. in their study related to annotation monetization. Rather than constant annotative tracking, Popscriptive will incorporate a relatively low number of annotations that only describe major scene changes and important visual elements in a conversational

sentence format that reflects the nature of the content. Due to simplified process and lower amount of required annotation information per video, Popscriptive will utilize user-designated keyframes without a complex editing tool, while still following the study's efficiency measures of constraining choice involving the structure of annotations. Popscriptive will also use an unpaid volunteer crowdsourced base, as was demonstrated in the StopFinder and VizWiz fashion-related applications, as timeliness is not a major concern for the Popscriptive service and it is more important to incur no financial cost. The system will incorporate crowdsourcing to build a public database of information in a similar manner to StopFinder, but enable participation through existing popular SNS, Twitter, rather than an isolated pool within the application. Twitter as a crowdsource base will also allow Popscriptive to scale rapidly, which is necessary for pacing with the high content upload rate of Youtube.

Notes

1 http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990217/

2 http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990217/

3 http://www.dcmp.org/

4 http://www.vocaleyes.co.uk/page.asp?section=166sectionTitle=Our+Services

5 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

6 http://harvardmagazine.com/2012/01/the-future-of-theater

7 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

8 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

9 http://ncam.wgbh.org/mopix/

10 http://www.datasatdigital.com/?option=com_contentview=articleid=49Itemid=33

11 http://www.doremilabs.com/products/cinema-products/fidelio/

12 http://pro.sony.com/bbsc/ssr/mkt-digitalcinema/resource.latest.bbsccms-assets-mkt-digicinema-latest-EntertainmentAccessGlasses.shtml

13 http://dcimovies.com

14 http://www.acb.org/adp/movies.html

15 http://www.acb.org/adp/dvdsalpha.html

16 http://www.narrativetv.com/

17 http://washear.org/

18 http://netflixproject.wordpress.com/

19 https://www.facebook.com/groups/324568721023571/335547796592330

20 http://www.parlamo.com

21 http://www.moviereading.com

22 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

23 http://www.youtube.com/yt/press/statistics.html

24 http://youtube-global.blogspot.jp/2009/05/zoinks-20-hours-of-video-uploaded-every_20.html

25 https://www.youtube.com/partners

26 http://www.hollywoodreporter.com/news/why-hunger-games-catching-fires-651660

27 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

28 https://www.google.com/accessibility/products/

29 https://www.youtube.com/yt/playbook/annotations.html

30 http://www.w3.org/WAI/

31 http://xiph.org/ogg/

32 http://www.w3.org/WAI/PF/HTML/wiki/Media_Accessibility_User_Requirements

33 http://www.miranda.com/Swift%20ADePT

34 http://ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie

35 http://imdc.ca/ourprojects/livedescribe

36 http://www.livedescribe.com/wiki/browse.php

37 http://www.inclusivemedia.ca/services/capscribe.shtml

38 http://www.vdrdc.org/research/dvx

39 http://www.mielelab.org/projects/dvx

40 http://youdescribe.ski.org/rel/

41 https://twitter.com/BerkeleyBlink

42 http://web.mit.edu/vondrick/vatic/

43 http://www.karimali.org/flowboost.htm

44 https://www.mturk.com/mturk/welcome

Chapter 3

Concept Development and Implementation

The initial goal of the research was to design a web-based product with the target user of a visually-impaired person. The time frame for the research extended for a one-year period from June 2013 to July 2014. In order to find a focus point for the research, exploratory fieldwork was conducted through volunteer activities in which vision-impaired people were main the participants. The purpose of the fieldwork was not to systematically discover a problem, but to achieve a deeper understanding of the daily lives of vision-impaired people. For example, one group conversation during an English lesson at a vocational skill development center revealed the vision-impaired students' difficulty reading braille during winter months due to the loss of feeling in the fingertips from the cold temperatures in Japan. Another student added that his sweaty palms made reading braille equally difficult in the summer months as well. This type of knowledge acquisition was the key foundation of the fieldwork and the project as a whole. The concept for Popscriptive was developed from the fieldwork findings, along with a review of the existing works and literature.

3.1. Exploratory Fieldwork

The author of the paper conducted exploratory fieldwork in three key locations. In all fieldwork situations, the author functioned as a participant or volunteer in

an immediate activity rather than as an purely outside observer with research purposes alone. The author was able to contribute to each fieldwork experience in some capacity as an English speaker. In this way, the author was able to provide some value to vision-impaired people at each fieldwork location to create a two-way exchange. Under the paradigm of participant observation established by Jospeh Howell, the researcher went through the steps of establishing rapport, going into the field, recording data, and analyzing data [20]. All initial exploratory fieldwork took place in Tokyo, Japan. The exact names of the locations will not be disclosed in the interest of privacy.

3.1.1 Fieldwork 1: Junior High School Mock Eiken Exam

The first location was a junior high school for the vision-impaired, where the author participated as a volunteer administrator of the spoken mock exam for English (EIKEN). Fieldwork was carried out on June 17, June 25, and October 25, 2013. Communication with the vision-impaired students was limited outside of the pre-approved testing script and evaluation period, but the author was able to experience the different available testing methods adapted for the vision-impaired exam takers. Although the test content could be accessed through braille, enlarged text, and a CCTV (closed-circuit television) system that enables magnification of the test material through a video camera that projects the test image onto a television screen, actual test content was not adapted or changed in consideration of the vision-impaired students. Testing instructions also lacked any adaptation, as well as testing format. For example, there was no additional time given for the reading test portion for students using CCTV or braille versions. Graphical content also remained unchanged, although an accompanying description was offered in the case of braille test takers who could not be expected to see the content in its original visual form. Corresponding questions were not altered to reflect the non-graphical description. Many braille test takers seemed to struggle with answering the graphic-based questions as they had to read a large amount of descriptive detail before they could filter out the information necessary to answer the questions.

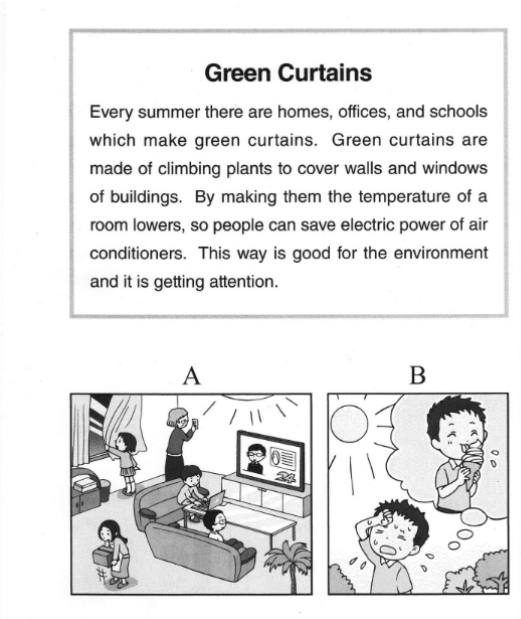

Figure 3.1: An example of a visual question from the Mock Eiken Exam.

3.1.2 Fieldwork 2: English Class at Vocational Development Center

The second location for fieldwork was a vocational skill development center where the researcher served as an English teacher for the beginners level Business English class. The class is held weekly on either Tuesday or Thursday, but volunteer teachers rotate the teaching schedule on a once-a-month basis. The author participated as a teacher for a one-year period from June 2013 to June 2014. Each class consisted of around 5 to 6 students. Students ranged from high school age to over 50 years old. Lessons consisted of free conversation in addition to teaching aloud from a textbook of 15 units designed around beginner's level business conversation, such as "What's the budget for the website?". As was found during the Mock Eiken Exam fieldwork, the textbook was not designed for vision-impaired students and the content was often visually-based. Teachers of the class adapted the content of the textbook each week by changing question formats during the lesson. For example, instead of fill-in-the-blank answers, the teacher would rephrase the problem into a spoken question format. Students sat at computer desks with one computer in front of each student and some students took notes using a word editing program on their computers using one side of in-ear headphones to hear the computer speech-to-text program read the typed notes aloud. Others chose to simply listen to the lesson. Display monitors were sometimes turned on but often

remained off. When turned on, the color of display monitors consisted of black and purple as the Windows OS was set in a specific contrast mode for better viewing for those students who still retained some level of usable vision. In addition, no mice devices were present in the classroom as all students used the keyboard to interact with the computer. Two younger students (high school aged and new university graduate) in the Thursday group used a refreshable braille note taking device. The only reference material available to the students was their own notes as all other parts of the lesson were spoken.

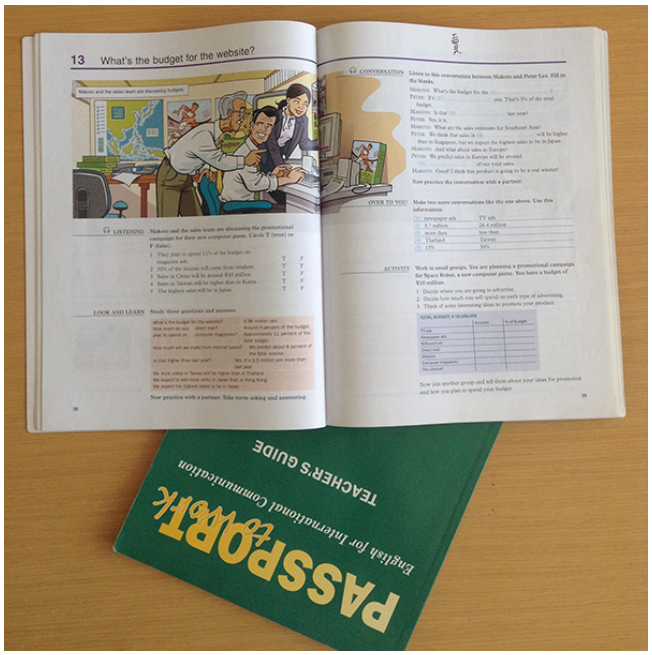

Figure 3.2: The English textbook used at the vocational skills development center for the vision-impaired is full of pictures and visual-dependent questions.

Students interest in learning English ranged, although many expressed an enjoyment of the class. Younger students seemed to show more interest in learning and volunteering to answer problems and make example sentences. Some students also expressed an interest in English language content, such as English language music or English language books. A few students showed a keen passion for trivia and knowledge gathering. Others had very specific hobbies, such as playing traditional Japanese drums. A couple students mentioned watching movies as their hobby. When questioned further, the students described how they traveled to a

specific cinema that provided audio description through headsets. Another student shared his enjoyment of American comedy shows such as The Simpsons and South Park, explaining that the humor style was very different from Japanese shows.

In addition to the monthly English class, the author also observed a computer skills class taught previous to the English class on Tuesdays. Students interacted with their computers using the keyboard and a set of in-ear headphones with one headphone left out of their ear in order to to hear the teachers instructions, as well as the audio information from the screen reading software. Unlike the English class, every student turned their computer display monitor on during the computer skill class in order for the class assistant, a sighted woman with no vision-impairment, to go around and check each students work. The teacher of the class was a vision-impaired man in his fifties who lost his vision later in life. He prepared the lesson documents previous to the class and distributed the files through the shared file system, having each student copy the files to their own personal directory. In addition to receiving feedback from the sighted classroom assistant, he also checked students work by having them unplug their headphones and listening to the screen reader, by sitting in the students seat and taking their headphones to listen to the screen reader, or, if their work was believed to be completed, having them print out their work which he would check using a magnifying glass held up to the printed paper. This last situation presented some difficulty because often the text was still too small for the teacher to see with the magnifying glass and he could not precisely check the words or formatting of the text. Students also struggled with checking with formatting, as well as remembering keyboard shortcut combinations and navigating menus — in particular, finding and remembering certain functions on the various tabs of the Microsoft Office Ribbon.

The computer skills program itself spans 6 months and includes the following topics of training:

Basic skills of using Microsoft Windows (total of 15 hours)
Introduction to JAWS screen reading program; how to use the JAWS cursor
Folder management (customization); display (customization)

Scaling options
Recording words to dictionary for writing in Japanese
Other

Basic skills for Microsoft Outlook (total of 20 hours)
Sending/receiving email; schedule management
Managing the address book
Managing jobs using the Task function
Memo-taking using the Memo function
Other

Basic skills for using the Internet (total of 20 hours)
Setting the place marker; accessing information quickly by using the shortcut keys

Basic skills for Microsoft Word (total of 60 hours)
Setting the page format, layout, and font style
Setting the even distribution in the layout
Setting the supplementary by Hiragana and insertion of greetings
Setting items (with numbers, alphabets, etc.)
Making the chart, adjusting the layout + Making the labels
Setting the printing process and the range of printing
Others

Basic skills for Microsoft Excel (total of 75 hours)
Page setup; using the AutoSum function
Inputting consecutive data; customizing display
Adding and deleting sheets, cells, rows, and columns
Editing the sheet name; moving data and extracting data
Manipulating data using the Pivot table function
Setting the input rules; password protection; inserting graphs
Processing data by using functions

Basic skills for Microsoft PowerPoint (total of 45 hours)

Making a presentation; manipulating the place marker
Adding, deleting, and changing the slide order
Inserting and editing text

Figure 3.3: A regular keyboard and a braille refreshable display.

As evidenced by the curriculum syllabus, a great deal of focus is placed on mastering Microsoft Office programs for job preparation. Skills for the Internet, in comparison, only received less than 10% of the total lesson time (20 hours vs. 215 hours of Microsoft Office-related lesson time). The Internet training unit also consists of only one module: Setting the place marker; accessing information quickly by using the shortcut keys. Other than basic browser navigation, the module does not deal with any specific site or content navigation or accessibility. In general, there is no emphasis on using the computer for entertainment or non-work purposes. Although this is not surprising considering that the program is for vocational skills development, the scope of Internet technology instruction is still markedly limited considering the importance of social media and other Internet services in the business realm, as well as the personal realm.

3.1.3 Fieldwork 3: Art Print Show

The third location for fieldwork was an annual art show event held by a volunteer organization for fundraising purposes. The volunteer organization activities in-

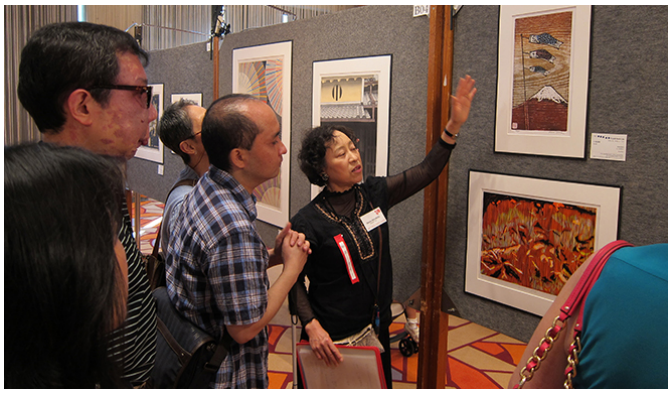

Figure 3.4: A group of vision-impaired visitors receiving a spoken description of the print.

clude running English language programs for the vision-impaired and awarding annual scholarships for vision-impaired students to study both domestically in Japan and abroad. The art show event is an annual event that first began in 1956. In 1996, the Hands on Art program was introduced in addition to print Show. Hands on Art presents a limited number of art prints selected from the current year's offerings that are recreated as tactile prints called "raised images". The organization created the raised images under guidance from the Japan Braille Library. Two particular prints were also adapted as relief works, which are 3D reproductions of the prints that are sculpted out of a plastic-like material. Visually-impaired attendees can tangibly experience the raised and relief images and participate in a tour by a volunteer who will give descriptive and background information about other paintings hanging in the show. In some cases, the participants also have the opportunity to hear information from the artists themselves.

The author participated in the art show in the capacity of a volunteer for the Hands on Art program for one day in October 2013. Volunteer duties consisted of escorting the vision-impaired attendees to and from the station to the venue, reading art information and passing braille information sheets to the attendees while the attendees explored the raised and relief images with their hands and fingers, and guiding attendees along the tour of the hanging art prints while explanations were spoken by another volunteer. There was a total of 20 visually-

Figure 3.5: Vision-impaired visitors to the art show feeling the tactile prints.

impaired participants over the entire course of the art show. As many of the vision-impaired attendees still retained a level of vision, some participants preferred the tour because they were introduced to a wider variety of art prints. These participants would often need to lean close the art print or view the art print from a particular angle in order to view the content of the paper. Larger pieces sometimes featured larger detail which were easier for the attendees to view, but also posed the problem of being difficult to take in as a whole view because of the large size.

3.1.4 Additional Fieldwork

The author participated in two separate social events for visually-impaired people on November 2, 2013 and March 29, 2014. These events were designed to be friendly gatherings where vision-impaired people could mingle with English-speaking Japanese and non-Japanese members of the above mentioned volunteer organization for English practice, as well as establishing communication between other vision-impaired people. In fact, many attendees mentioned their motivation for participating in the event was to meet other vision-impaired people, although others complained that there was not enough actual English practice due to this secondary motivation.

The first gathering took place in a church and featured a concert by a blind

musician, as well as group singing. As vision-impaired participants could not read the lyrics of the songs from the provided paper, sighted participants read out the lyrics first and vision-impaired participants repeated the lyrics until the words were satisfactorily committed to memory. After the church concert, vision-impaired and sighted participants mingled and spoke English during lunchtime.

The second gathering took place in a rented meeting room where over 30 attendees assembled to listen to an assortment of vision-impaired guest speakers who shared their different experiences. One guest speaker talked about his recent trip around the world where he visited over 20 countries. He talked about the difficulties he experienced — as a foreigner, as a sole traveler, and as a vision-impaired person, along with the warmth he found in people around the world who made an effort to assist him in his travels. One anecdote he told was when he was on the subway in the United States and asked a man for assistance. Although the man ignored him, other passengers on the train immediately offered their help. He encouraged the other vision-impaired attendees to travel the world and create their own experiences. Another two speakers, both teachers at public schools in Japan, shared about their recent trip to Korea to meet with other public school teachers for international exchange. When they initially traveled to Korea, they did not have any particular contacts — but they were somehow able to find other vision-impaired teachers and plan a meeting. Afterwards, they shared some trivia about Korea, having other attendees guess the answers by clapping for their selection out of the multiple choices offered. While the speakers told their stories, they used note-taking tools for cues, particularly considering that they made their speeches in English rather than in their native tongue of Japanese. Some speakers used a refreshable braille display note-taker and read the braille simultaneously to speaking to the audience. Other speakers used audio notes, by keeping a single in-ear earphone in one of their ears and listening to their cues. As a final activity for the gathering, the room was split into different groups and the different groups competed in an animal-sound guessing game, as there is a significant difference between animal sounds in Japanese and English. Although the game was created with vision-impaired players in mind, the game cards were highly textual (visual text and braille) and required the volunteer to descriptively convey the meaning of the different cards to the users who could not read the available braille.

3.1.5 Content as a Common Theme

A common theme found during all exploratory fieldwork was the issue of accessing and adapting content — whether in the form of testing materials, textbooks, software, or art prints. In some cases, accessibility to content is offered through a human volunteer, a text-to-speech screen reading technology, or through scaling and/or printing techniques. Although some content had been adapted for better accessibility and understanding, as seen in the case of the Art Print show, the adaptation was very constrained due to time and cost intensive nature of producing the "raised images". Out of the 188 different art prints showcased, only 5 were recreated as "raised images" or relief works. In addition, there is some debate as to have true the adaption is to the original and how meaningful the adaption is to the vision-impaired user. For example, the "raised images" provided a novel experience for the vision-impaired attendees of the art show, but attendees still relied heavily on explanations by sighted volunteers in order to understand what the individual parts of the raised images were expressing. This descriptive information included a written script combined with ad-libbing by the volunteer for color and other details. In addition both forms of adaptation, raised and relief images, featured simplified versions of the original art print as it was deemed impossible to include all detail. For example, an etching by Mayumi Someya entitled "Further On" featured nine snowmen in its original print form, but only four snowmen in the adapted form. Another work by Kevin Lee Clark entitled "Koinobori" was originally painted on oak with a strong wooden grain pattern (Figure 3.4). In the raised adaptation, this wooden grain is lost. Although there is alterations to the original works, the response to the Hands on Art event has been continuously positive over the years with participants expressing a sense of enjoyment in the new experience of feeling the art. The vision-impaired visitors demonstrate a sincere desire for access to the art print content through either tactile reproductions or spoken descriptions.

In some cases, human interaction was necessary for ad hoc adaptation, as seen in the adaptation of the English mock exam and textbooks. In these particular situations, volunteer human labor was a more cost efficient alternative to creating or purchasing specialized material for the vision-impaired user. In many cases, it is not a viable option to provide specialized or customized content due to financial or

time-related constraints, or simple due to unavailability. Concurrent to adaptation of the content, the human volunteer also filters content based on their personal judgement of what is critical or non-critical for understanding, as well as what is meaningful or not meaningful for the vision-impaired subject. The human volunteer may add their own ad-libbing or explanations, as well as their own personality into the adaptation which can serve to benefit or obstruct the vision-impaired target audience in their comprehension of the content.

Figure 3.6: The interaction between a vision-impaired person, the content, and a sighted volunteers.

3.1.6 In-depth Interview About Technology Usage

Due to the relatively limited exposure to vision-impaired technology usage during the exploratory fieldwork, an interview was conducted with one of the attendees of the art print show about his use of other technologies. The interviewee was a Japanese male in his early twenties who is currently attending university in Tokyo, Japan. He began suffering from vision impairment around age 13, with peripheral vision loss due to retinal disease. He still retains a level of vision and wears strong prescription eyeglasses. His goal is to work in advocation of accessibility and equality for vision-impaired persons. The interview was conducted in English as the interviewee understands and speaks English with a high level of fluency.

Q: Could you tell me about your current technology usage (for example: Do you use a smartphone)?

Now I don't use a smartphone because it is inconvenient for me due to its small

words. I understand its usefulness, but I feel difficulty using it. If there are some apps that are easy for people with visual impairment to use, I want to have a smartphone.

I use iPad and often use Skype, Facebook, Twitter, news apps, and so on. It's easy for me to see, not only because the words are bigger but also because the screen light is clear enough. For my eyes, clearness of screen is important because the darker it gets, the less I can see due to my disease in my eyes. Even if words are so big, I can't read them at all without light. I've heard that some people with visual impairment care about the color of words.

When people with visual impairment use computers, many use a software called screen reader, that reads loudly sentences on screen. Though I don't use it because I can read sentences directly. I don't know whether screen reader is available in smartphone or tablet. Anyway, it is better to make apps easy to use with screen reader, I think.

Even though I am very busy with studying, I can't get rid of my television because I want to watch One Piece, a Japanese popular anime.

On my cellphone, other than e-mail and calling functions, I often use EZ navi walk, which is a map search and transfer information app, and LISMO, which is a music app. I use au cellphone. I don't use web by using cellphone because words are too small to read.

Q: My favorite app right now is the Nameko[1] *app. Have you heard of it?*

I know Nameko. Many of my friends enjoy it. I remember that when my friends played it, I heard strange sound from their cellphones.

Q: For map search applications, do you prefer to look at the map itself or do you instructions (such as "Go straight to XX Street, then turn right on YY Street)?

I prefer instructions because it takes time to read maps and reading maps makes me tired. When I read something, I need to use zoom- up machine to read, and it is troublesome taking it out of my bag, looking for pages I want to read, and understanding where to go. Hearing is much easier.

Q: What about online video SNS? Do you use Youtube or Niconico Douga?

I often use YouTube, but I've never used Niconico Douga. I mainly watch music PV, especially Yuzu's PV, who is my favorite singer.

Q: Do you ever like watching amateur content (versus professionally created content like music videos)?

I don't often watch amateur content, but sometimes I watch when I have time.

From the interview with the vision-impaired young adult, along with the observations during the exploratory fieldwork, the author found that many of the vision-impaired students and participants of all ages used computers and iPads, but still preferred traditional style cellular phones to smartphones. People enjoyed all forms of entertainment, including highly produced video content such as movies and television shows. In addition, as stated in the interview, Youtube was used primarily for music rather than for user-generated content. This trend does not differ from Youtubes own reported video view statistics which show that 9 out of the 10 most watched views are professional music video content.[2]

One important point raised during another interview with a completely blind Youtube user was the importance of access to culture. "We like culture. So it's not that we want to watch the animal videos, it's just for curiosity's sake. It's the cultural literacy. I just want to know what's popular, what my friends are watching...cats... [...] For example, with Gangnam Style, I've heard the song — but I also know there's something more, like a dance. But I don't know the dance." When the researcher described PSY as a middle-aged Korean man who danced as if he was riding a horse, the interviewee laughed and displayed appreciation for the new descriptive knowledge. Another interview with a man with partial blindness demonstrated his clear dismay with the current condition of accessibility to Youtube. "I picture, I don't know why I picture it like this, but people sitting at a call center. Sitting at, like, a call center in lines, verbally describing video. That should be bunch a people's jobs to do that. Why hasn't Google done that?

They work on accessibility." Accessibility to content can be seen as accessibility to culture and it is necessary for all people, regardless of their level of vision, to be able to encounter and connect with culture in the same manner as a sighted peer.

3.2. Concept Development for Popscription

After data from the exploratory fieldwork and interviews was collected, the researcher identified content accessibility as a core focus for the project. Considering the initial goal of creating a web-based tool for vision-impaired people, the author decided to concentrate specifically on web content. User-generated online video content accessibility, in particular, presented a very challenging, and perhaps, neglected, situation with great potential for accessibility and adaptation due the sheer volume of existing content, rate of new content creation, along with its rapidly expanding role in contemporary culture and expression. It is not feasible for videos to be professionally adapted through audio description in the manner of high-budget Hollywood films. At the same time, it is not practical to expect real-life human volunteers to describe video content to a vision-impaired person on a daily basis. Furthermore, dependence on another individual can have a negative impact on the self-reliance and freedom of exploration that the Internet provides as a tool for the vision-impaired person. A solution would need to be unobtrusive to the Youtube experience while providing descriptive annotations suitable for the web video format. Although data would still need to be provided by sighted volunteers, it could be provided collectively in a database for free access which would eliminate dependence on the actual presence of the volunteer for content information.

Although the audio description provided for high-cost production films and television involves a professional narrator reading descriptions timed precisely around original dialogue and soundtrack, the resource-intensive requirements for this traditional style of audio description is impractical for the medium of online video. The design of a new format of audio description was necessary to match the specific characteristics of online video. In particular, the brevity of web video makes it difficult to audio describe. In many cases, the description of the video

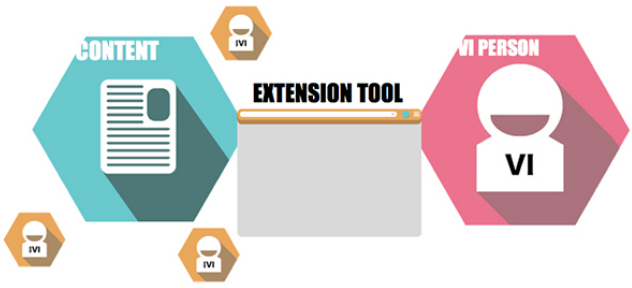

Figure 3.7: The proposed model for interaction between a vision-impaired person, the content, and sighted contributors.

would take longer than the video itself. An envisioned solution to this problem was to pause the video in order to deliver the description without overlapping with existing dialogue or running overtime.

For the format of user-created web video, the daily mass of new content is also a key factor for consideration. While the production process can take weeks, months, or even years for a high budget film or television show, web content is produced with little to no budget and in very short periods of time. This results in an exceedingly large quantity of new content being uploaded everyday by users worldwide. In order to match this prolific content growth, it is necessary to design an efficient process for providing audio annotations. With the perpetual outpouring of user-created content, it is logical to create a parallel system that depends on users for user-created annotations. In order to attract users to supply annotative information, is integral to design a simple process without technical or time barriers that will discourage people from contributing. One such potential barrier would be the necessity of voice recording. To require contributors to not only create descriptive information, but also provide an audio recording of acceptable quality could be too demanding and may significantly limit the inflow of annotation information. Therefore, the proposed solution will utilize speech-to-text technology that will create synthetic voice descriptions out of user-created text annotations to eliminate the task of recording for the human contributor.

It was also crucial to create a solution that would utilize the existing content, community, and structure of Youtube. Creating another platform would not only be redundant and inefficient — it would also result in isolating the vision-impaired user from the mainstream Internet, which would only compound the problem of

access to content and culture. The solution would need to be completely integrated with Youtube, so as the user would not feel any significant change in the overall site experience. In order to achieve this goal, the researcher built a browser extension which would act as a supplement to Youtube rather than a replacement or alternative. The extension works invisibly in the background of the browser, only becoming active when the user visits a supported webpage, in this case, an individual video page. As soon as the video page loads, the user can press the keyboard shortcut to access the extension and start playing the video with descriptive annotations. The user can interact normally with Youtube, but also receive the annotation information.

3.3. Prototyping Method for Popscriptive

In order to build Popscriptive as an effective tool for the target vision-impaired user, it was very important to consider the user from the very initial stages of design through a prototyping method. A study by Phillips et al. demonstrated that 29.3% of all assistive devices were completely abandoned [34]. In order to prevent user abandonment, the research found that consideration of user opinion in selection was found to be one of the key necessary factors to prevent abandonment. As stated by Winograd et al. in *From programming environments to environments for designing*, it is critical that prototypes have a "feeling of roughness" in order to attain substantial user opinion and feedback [52]. "A highly polished prototype — even if it only a first attempt at the functionality and interface structure — fosters a sense of finality that tends to inspire only suggestions for minor improvements and further visual niceties [52]." The purpose of prototyping is to encourage communication and Popscriptive followed this philosophy when testing the initial features and concept with users.

Although Popscriptives prototype lacks the extreme rough sketch style encouraged by Bill Buxton in his book, *Sketching User Experiences*, it follows his principal attributes of being quick, timely, inexpensive, disposable, plentiful, clear in vocabulary, distinct in gesture, minimal in detail, appropriately refined, suggestive and explorative rather than confirmative, and ambiguous [9]. In addition, the prototype balances between "too little fidelity" and "too much fidelity" which

renders an idea to perceived as already completed.

Popscriptive uses Popcorn script[3], a Javascript library that enables video, audio, and other media to interact with arbitrary elements of a webpage either in a control or be controlled relationship. The HTML5 media framework is an open source project by Mozilla. The library makes it simple to interact with the media elements without comprehensive knowledge of HTML of Javascript. In this way, Popcorn script enables a rough sketch style while still being built with functioning code to create a working prototype. The extension refers to a PHP web application which connects with the MySQL database to print the annotation information, or lack thereof, back into the extension pop-up window. The annotation information is originally stored by another PHP web application which utilizes the Twitter API to scrape properly formatted tweets into the MySQL database for later retrieval.

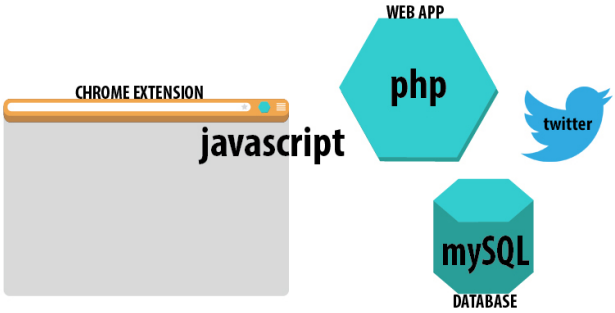

Figure 3.8: Popscriptive consists of components built with Javascript, HTML/CSS, and PHP.

3.3.1 One System, Two Models

Popscriptive as a system will consist of two different models for input and output that combine together for the ultimate goal of providing casual annotation for Youtube videos to create a better experience for the visually-impaired user. As Popscription has two unique models for two different target user groups, the prototype study was split into two separate studies.

The first model is based on the output side of the Popscription: the presentation of user-generated annotations of visual detail for Youtube videos. The target user group for the output side of the proposed system is vision-impaired Youtube users who wish to access and consume video content. Vision-impaired

users will access the Popscriptive extension through the toolbar of the Google Chrome browser or the keyboard shortcut. The extension will only be available when the user is on a Youtube.com individual video page URL. After clicking the extension, the video's unique identification number will be searched for in the SQL database hosted on the independent Popscriptive server. If the video's ID number can be found, the matching stored annotation information will be called. This information will be presented in the pop-up of the extension along with the original video content from Youtube. If no information can be found, an error message will appear in the extension pop-up and the user will have the option to request for an annotation via tweet by the Popscriptive Twitter account.

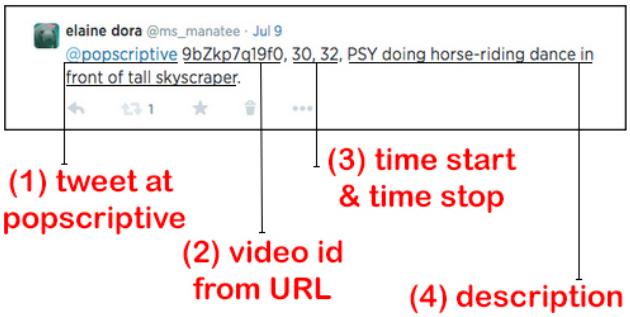

Figure 3.9: Tweet format for annotation input.

The second model for Popscription is on the input side using the Twitter API to collect designated tweets from volunteers who create the annotations along with an appropriate time stamp and Youtube video ID number. The target user for the input model is sighted Youtube and Twitter users who have an interest in volunteering to assist vision-impaired people or are Youtube content authors who wish to make their content more accessible and/or reach a wider audience of viewers. An interested volunteer or content author can tweet annotation information to the designated Popscriptive Twitter account using their own Twitter account. The tweet must be formatted with the Youtube video ID and the timestamp for when the annotation should begin to be presented. The ending time for the annotation presentation will be designated by the stop annotation time stamp. Due to the Twitter-imposed limitation of 140 characters per Tweet, annotation providers will be forced to write their descriptions with desirable brevity.

3.3.2 Three Potential Delivery Methods

The first output model user study consists of three different prototypes with the same basic code structure using the Popcorn script library but with differing annotation delivery methods. All prototypes are HTML pages that are accessed through the Popscriptive Chrome Extension in the Chrome browser which only becomes active when the user is on a Youtube page. However, the extension interaction was only tested once during the user study as it was unnecessary to duplicate the interaction to test the three annotation delivery types and would only detract from the focus.

The first delivery method consists of annotation delivered automatically when the video reaches a specified time stamp which has annotation information available. In this case, the user cannot request or ignore the annotation, although they can cut the annotation short by pressing the enter key or play button on the embedded player to resume the video play. However, at the present, the audio will continue to be read out even as the video play resumes. The user must press the enter key in order to resume the video play each time after the description is finished.

The second delivery type consists of annotation delivered at the user's request through pausing the video play by pressing the enter key. The timestamp of the users request will be searched against the annotation database to see if there is a an annotation available for the specified time frame. If an annotation is available, it will be presented to the user. This method gives the user the freedom to request an annotation only when necessary, for example, in the instance that the user feels some uncertainty about the audio meaning of the video and wishes to see if video cues will clarify the meaning. The user can watch the video with less interruptions and more control over annotation information. In the same manner as the first delivery method, the user must press the enter key in order resume the video play after the annotation is delivered.

The third delivery type consists of an alert sound that notifies the user when a new annotation is available for the particular time range of the video. The user can then chose whether they are interested in hearing or reading the annotation. If they are interested in the annotation, they will pause the video by pressing the enter key and then press the enter key again to resume the video play after

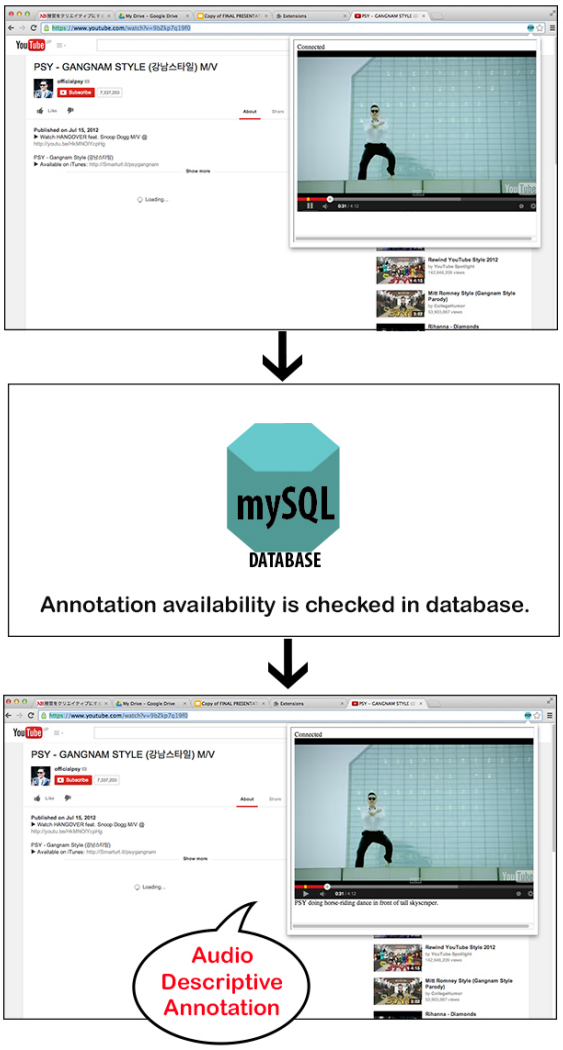

Figure 3.10: Prototype 1: Video will automatically pause for annotation descriptions to be read aloud by synthetic speech.

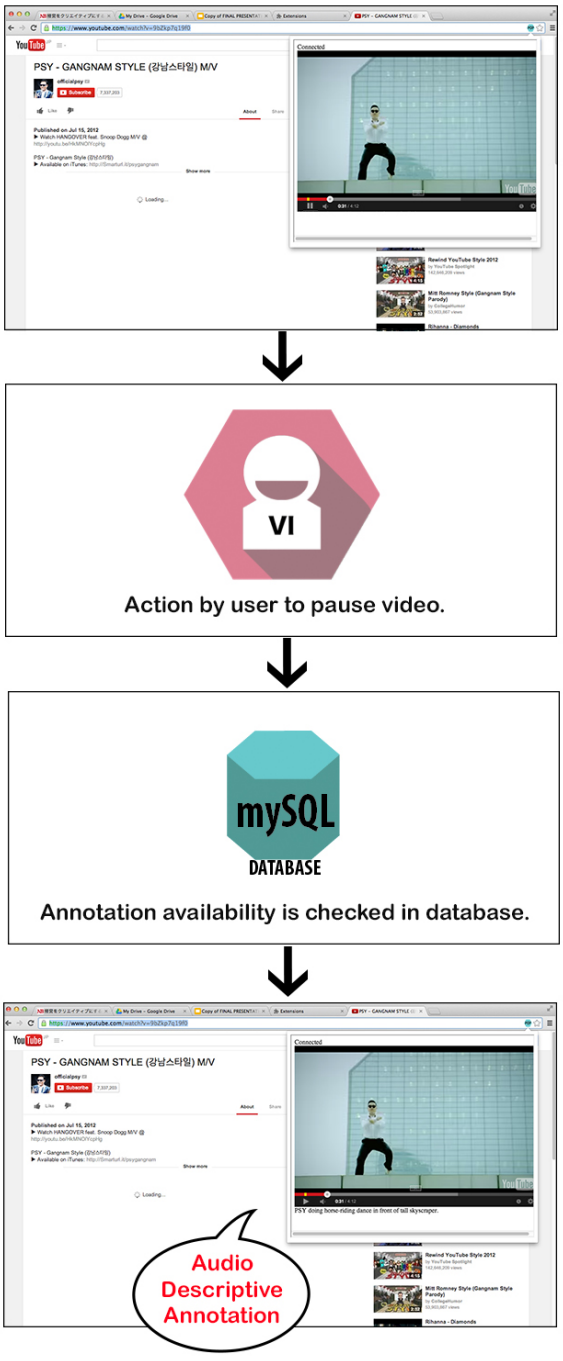

Figure 3.11: Prototype 2: Requires user action to pause the video for on-demand annotations.

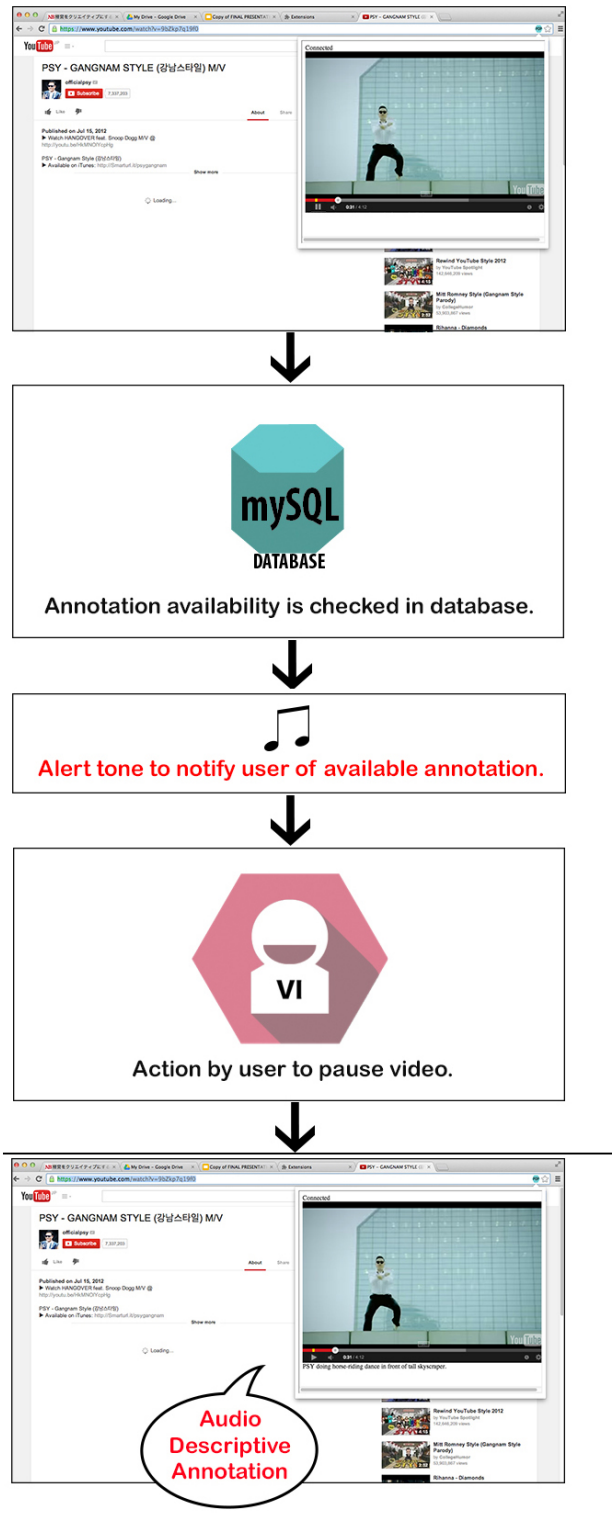

Figure 3.12: Prototype 3: Includes audio alert notifying users when annotation is available for a particular timestamp.

they are finished the annotation. This option gives the user the power to decide whether they wish to receive the annotation information, similar to the second prototype — except the user is informed whether information exists or does not exist before they make their request.

For the user study of the three various prototypes, participants were directed to watch specific videos with annotations prepared prior to the test. In an actual use setting, the user would be able to visit any video of interest on Youtube and check to see whether annotations were available in the database. The user was asked to watch the same video four different times. The first time was without the Popscriptive tool and the following three times with different versions of the prototype delivery method. While the participant interacted with the video, they were encouraged to perform a concurrent think-aloud to make comments and express any feedback about the tool or content. In addition, a pre- and post-interview was conducted to collect background information about the user and their video watching habits, ask for specific response to individual aspects of each prototype, and evaluate the relative effectiveness of the system as a whole. Key questions focused on whether the annotations were distractive to the content, if the vision-impaired user felt they could understand the characters and story more distinctly, and what content the vision-impaired user would like to use the tool for.

The second input model user study involved sighted Youtube users who have some experience using Twitter. The users were asked to watch a short Youtube video of their choice and then asked about their willingness or motivation to create annotations for other vision-impaired users. They were then asked to identify scenes that needed annotation and to write annotations with the 140 character structure. The participant had the option to tweet directly at the Popscriptive account from their own personal Twitter account or from a sample account provided for the study. The participants underwent pre- and post-interviews to collect background information about their Twitter usage and interest in Youtube and providing annotations for others. They were also asked to evaluate the process of annotation contribution to see what encouraged or deterred them from the given task.

3.4. Popscriptive Implementation

Based on the pop-up nature of audio descriptions provided by the system, the name Popscriptive (Pop + Descriptive) was devised. The system consists of a browser extension and a privately-hosted database. The extension will be downloadable through the Chrome Web Store under the Accessibility category at no cost. For the purpose of the user test, the prototypes of the extension were uploaded to a local Chrome browser through the Developer Mode which allows for the addition of unpacked extensions for testing. Once the extension is installed and enabled in the browser, the extension icon appears within the URL input space at the top of a browser. As a 'Page Action' type of extension, the extension will only become available when a user visits a designated URL — in this case, any Youtube.com individual video page. When a user accesses the extension, their current Youtube video is automatically redisplayed via the pop-up of the extension. Within the pop-up, text annotations appear under the video. These text annotations can be read aloud through speech-to-text technology or haptically through a refreshable braille display. There is also the option of enlarging the text for those users retaining a level of vision who would prefer to read the text.

To conduct a user study with the prototypes of Popscriptive, particularly focusing on the audio annotation delivery method, three versions of the extension were uploaded into the Chrome Browser through Developer Mode. Each version of the prototype uses the speech-to-text feature available in HTML5. The extension itself is built with HTML, CSS, and Javascript that removes the original player from the Youtube page and accesses a PHP file hosted on an independent server. The PHP file incorporates Mozilla's Popcorn JavaScript library to allow the Youtube video element to dynamically control other elements on the page within the extension. It also communicates with the MySQL database on the same server to access the stored annotation information.This enables the timestamp of a Youtube video to be a cue for action on the page – in this case, annotations to be retrieved from the database, added to the page, read by the speech-to-text function, and removed from the page.

Although the research focuses heavily on the vision-impaired user of online video platforms, due to the crowdsourced base of information necessary for the Popscriptive system to function, it was necessary to study both the vision-impaired

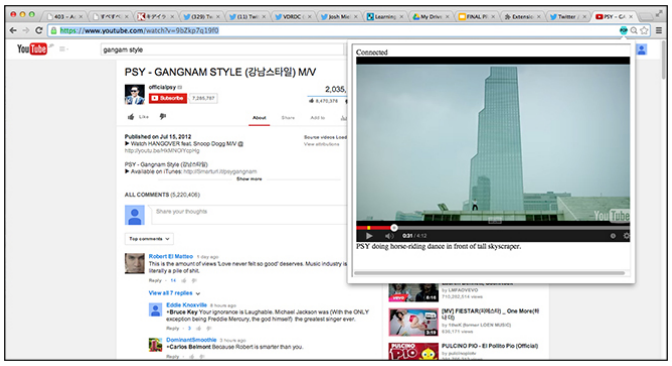

Figure 3.13: Popscriptive Extension in use on Youtube video page.

and sighted users on either sides of the input and output process. The prototype will be adjusted based on the feedback from both types of users. The results of the user study and evaluation, as well as the general knowledge gained, will be discussed in detail in Chapter 4: User Evaluation and Chapter 5: Conclusion.

Notes

1 Nameko is a mushroom farming game for iOS and Android where users grow mushrooms and try to collect different types of mushrooms. Nameko is a type of Japanese mushroom. The game is produced by Beeworks Games and achieved explosive popularity after it was released in 2011.

2 http://www.thedailybeast.com/articles/2013/04/23/youtube-s-10-most-watched-videos-ever.html

3 http://popcornjs.org/

Chapter 4

Evaluation of Research

4.1. Setting of the User Study

The user study consisted of in-depth interviews with 5 participants with various levels of vision-impairment. Interviews spanned a one-hour period and consisted of one-on-one interaction between the researcher and interviewee. In one case, two interviewees were questioned at the same time due to their status as a married couple. The location of the interview varied from the school of the researcher to the individual homes of the interviewees. Participants were first asked general questions about their background, general Internet experience, and their current web video experience. Afterwards, the participant tried Popscriptive system in its varying prototype forms and were interviewed.

A comedy skit video from Youtube was selected as the test content. The video has over 200,000 views and spans 2 minutes and 59 seconds. The video was specifically chosen for the humorous nature of the content and the necessity of visual cues in order to fully understand the humor. The same video was shown three times, each time with a different version of the prototype. Participants were asked for their general feedback on each prototype towards the end of each demonstration. At the end of all three prototype demonstrations, the participant was asked to give a full evaluation of the three prototypes and to chose the prototype they personally preferred, if there was one. It was also possible for participants to chose a mix of the three prototypes.

Table 4.1: Background information about user study participants.

Age	Gender	Country (Nationality)	Vision-Impairment
40s	Male	Japan (Sweden)	Low (Severe cataracts in both eyes)
30s	Male	USA (USA)	Med (NLP in one eye, legally blind)
20s	Male	USA (USA)	High (NLP in both eyes)
20s	Female	USA (USA)	High (NLP in both eyes)
20s	Male	Japan (Japan)	High (legally blind in both eyes)

4.1.1 Profile of Participants

The user test included 5 visual-impaired participants with a level vision to qualify them under the previously described definition of having significant difficulty seeing even with the application of corrective lenses. As to best reflect the varying degrees of visual impairment throughout the population, participants with varying levels of visual impairment were purposely selected. Level of vision-impairment ranged from severe cataracts to complete blindness with no light perception. Correspondingly, all study participants used various means for accessibility to visual interfaces and content. The participants aged ranged from the early-twenties to mid-forties. Of the 5 participants, 4 were male and 1 was female. The country of origin of participants included Sweden, the United States, and Japan. Interviews were conducted in Japan and in the United States. All interviews were conducted in English. Three participants received monetary compensation for their time. One participant refused compensation and another participant was not offered compensation.

4.2. Methodology

A qualitative method was selected as this study was specifically concerned with meaning. As described by Spencer et al. in *Quality in Qualitative Evaluation: a framework for assessing research evidence*, key features of a qualitative method include attention to subjective meanings for participants and the perspective of the participants, as well as the importance of context [45]. The intention of the user study was to attain a detailed response from the vision-impaired participants to annotation delivery methods within the context of Youtube video content.

As a qualitative study, answers were not measured on a scale or with a numbering system. In order to receive the most detailed responses possible, questions related to the evaluation of the three research goals were asked in a open-ended manner which allowed the participant to expand their individual, subjective opinions about the delivery design and took into the account the flexibility required for qualitative research. Furthermore, the evaluation followed the alternative quality criteria approach of Social constructivism which emphasizes subjectivity, multiple perspectives, reflexibility, and particularity [45].

The goal of the user study were as follows:

(1) To qualitatively measure the changed meaning of content with the addition of the Popscriptive tool.

Measurement for this goal in this particular user study was based on a true/false question statement to discern whether the content of the visual cue dependent sample video could be understood with little to no confusion or doubt.

(2) To qualitatively assess the technical quality and user interaction of the Popscriptive tool as a annotation provider.

The provision of three different prototypes for delivery method supported this second research goal. Considering the three prototypes, questions were asked to determine the basic technical usability success of the tool. A method of data collection that borrows from the SEM-CPU methodology proposed by Lee et al. for the systematic evaluation of cell phone user interfaces with consideration of accessibility issues due to disability was used [25] . The SEM-CPU methodology was adapted for the user study with the removal of quantitative aspects due to the low number of participants and the aim of the study to follow a social constructivist qualitative model. In addition, the evaluation methodology of the study used two empirical methods, rather than the original five of the SEM-CPU methodology, including concurrent think-aloud and post-task interview. Concurrent think-aloud was selected to replace the retrospective think-aloud method

chosen by SEM-CPU. Although the reasoning for retrospective think-aloud under the SEM-CPU methodology is supported by previous studies that demonstrate the disruptive nature of concurrent protocols, it was not appropriate or feasible to ask vision-impaired participants to watch a video recording of their task performance as the retrospective method requires [25]. In addition, the methodology was modified to fit the simplified user interface of the extension tool verses a cell phone user interface.

In the post-task interview, two key questions were asked to measure the technical effectiveness and usability of the annotations provided by the extension prototype and various delivery methods:

1) How was the audio quality of the synthetic annotations?
2) Was the delivery of the annotations disruptive or annoying?

(3) To gain insight about future design considerations for audio description tools like Popscriptive.

The open-ended style of questioning in the post-task interview combined with the data collected from the concurrent think-aloud provided design considerations that can be incorporated into later version of Popscriptive or be used by other audio description research and development.

4.3. Procedure

Each user study can be broken into three sequential segments. The first segment is a pre-interview, where the participant is asked about basic background regarding their vision-impairment and usage of technology. In the second segment, the user interacts with the Popscriptive prototypes. The same order of prototype presentation was used during each user study. The first prototype featured the delivery method that automatically pauses and delivers available annotation based on the time stamp information. The second prototype featured the delivery method that requires the user's active pausing to retrieve annotation information. If no annotation information exists in the database for a particular time stamp, a message

stating "No annotation available," is heard instead. The third prototype featured the delivery method that alerts the user when an annotation is available with a brief tone. The user then must actively pause the video if they wish to receive the descriptive information. During the presentation of all three prototypes, the participant was encouraged to engage in concurrent think-aloud about their interaction with each prototype model. In the third segment, after the prototypes have been shown to the participant, a post-task interview was conducted. The post-task interview consisted of semi-structured questions which aim to measure the extent to which the the first and second goals set previously were accomplished. In addition, the semi-structured format allows for free response for open learning in order to accomplish the third goal of gaining insight for future design considerations.

4.4. Main User Study

4.4.1 Participant 1

The first interview was conducted on June 5th, 2014 in a meeting room at the Hiyoshi Campus of Keio University. The participant was a male in his forties from Sweden who has been living in Japan for over 10 years. Six months prior, his vision deteriorated very rapidly due to the development of cataracts in both eyes. Although the condition is operable, he had been unable to undergo the procedure due to personal issues. As an avid user of the technology, the sudden deterioration in his eyesight forced him to try accessibility options, mainly contrast and sizing-related, to continue his normal lifestyle. He uses Windows 7 and the Mozilla Firefox browser with all accessibility options for enlarged fonts and contrast adjustments. He also used keyboard shortcuts to the best of his ability, but still struggled with familiarity. The participant provided the perspective of a person who very recently began struggling with vision-impairment and a person who still maintained a relatively high level of vision.

Although the participant was an avid user of Youtube before his vision problems, he currently found the site very difficult to navigate and the size and contrast of videos to be too difficult to see. He needed to move very close to the screen to

see the content, but mentioned that, "video is easier than reality actually. There's more light into the eye." Although the participant rated his past user experience with Youtube as an 8 on a scale to 10, he rated his current experience with his vision-impairment to be a 3 out of 10. In general, the participant uses Youtube for entertainment rather than educational or instructional purposes.

As the participant still retained a relatively high level of vision, he still preferred to use his eyesight rather than rely on audio description. However, as he could not see the visual details of the video, he still found the annotations necessary. He recommended enlarging the annotations and player buttons to "at least 5 times the size", saying that he could not see either in the any of the current prototypes.

When questioned as to what prototype he found most effective, the participant answered, "I think it's very subjective based on your level of impairment. Because I see two fuzzy guys but I don't know what they are doing. If you are more impaired, you probably want the automatic pause [Prototype 1] so that you can follow what is going on.

If you are less impaired than I am now, you probably want to be able to choose. It gives you support for understanding the ongoing video. Obviously that is less distracting than pausing the video."

The participant chose the second prototype as being the most useful because he only needs the annotation at specific points in the video where he cannot see small details, such as what was being held in the hands of the actors in the comedy skit. As a hybrid between the second and third prototype, he suggested adding a visual pointer or overlay to illustrate where an annotation was available verses the audio alert.

In all three prototypes, the participant evaluated the prototype as being successful in providing deeper meaning to the content. In response to the technical evaluation questions, the participant found the synthetic audio to be clear and understandable although he expressed some dissatisfaction with the max volume.

4.4.2 Participant 2

The second interview was conducted on June 27, 2014 over the telephone. The participant was a male in his thirties from the United States. He has been vision-

impaired since birth due to the condition of Retinopathy of Prematurity. He is classified as legally blind in both eyes. He has no light perception (NLP) in his right eye. He describes the vision of his left eye as "20 over not good".

The participant uses computers on a daily basis, stating that he is "looking at a screen all day." For Internet browsing, he uses Internet Explorer at work, but at home, he also uses Mozilla Firefox. He uses a combination of screenreader assistive technology, a text zooming program, and keyboard shortcuts, but still relies on his remaining vision "a little too much." He states that, "I wish I was better at not using the mouse". His leisure activities include sports and event planning. He does not watch movies or television on a regular basis.

The participant was an active user of Youtube for music and entertainment. The participant expressed an eager interest in the annotation system but had mixed opinions. He wanted the system to be closer to the existing DVS (Descriptive Video Service) in movies with no pausing, but also stated that he could understand a different system for the content style of Youtube. When shown the three prototypes, the participant initially believed he would prefer the 2nd prototype due to the optionality of the annotations. However, after actually trying the prototype, he found it to be "not very user friendly" because of the frequent unavailability of messages which resulted in the "No Annotation Available" default message. He also commented, "for people who are totally blind, they would be guessing because they can't see whats going on the screen. I can see a little so I can tell when something is happening, but people who can't see it, someone who is totally blind would not know to press the pause button."

The participant found the first version of the prototype to be very "straight" but showed the most interest in third prototype. He stated, "I think either you want [description], or you don't, going into it. But I'm changing my mind here as I'm talking to you, sorry. It's nice to know whats going on a little bit. It's nice to have a little bit of background of whats going on.

I don't know if the third one with the chime, it's gotta be there 100% or it's not. So it might be too much for a movie. In the case of three minute thing like I just watched verses a movie — I think that's different verses a movie. I would know whether I wanted DVS right off the top.

For a short video, if it's fast paced maybe I want it on all the time. If not, I

may want to take breaks. I personally have some usable vision so I can sit there and look and it."

The participant rated all three prototypes as giving more meaning to the web video. Rather than choosing a prototype he preferred, the participant stated he would want the option to chose between the different delivery types based on the video pacing or duration. He elaborated, stating, "I tell people I don't like action movies because they're real fast and I can't see the action. I don't watch movies often because I can't tell the difference between two people. I'm just just going off voices most of the time."

The participant said he would like to use the annotation system for exercise videos, explaining that currently "it's not exactly descriptive of what motions are involved other than the names of the exercises themselves." He continued, stating that, "it moves so fast that you can't. It will take the whole minute to describe what's going on presumably."

In response to the technical evaluation questions and the issue of live verses synthesized voice, the participant stated his preference for live voice, but was also willing to concede live for synthesization in the case of the Youtube video format as presented in the prototypes. As annotations for Youtube videos would be crowdsourced, he stated his preference for the guaranteed quality of synthesized audio rather than a varying quality of live voice description.

4.4.3 Participant 3 and 4

The third and fourth participants were a married couple in their late-twenties. As the two lived together, the interview was conducted simultaneously with both participants in their home in California, U.S.A on July 2, 2014. Both participants were blind with no light perception. In fact, the participants informed the researcher that it was unnecessary for the two of them to turn on the lights in their house. Describing their own level of vision, one of the participants commented that her vision was "nonexistent". The male participant had retinoblatoma at 18 months of age which had led to his blindness while the female participant "had too many conditions to name when I was born". Both participants were using multiple screenreading programs when accessing their computers and smartphones, including JAWS, PVDA, System Access, Windows Eyes, and Voiceover. Both

participants were also avid Internet users, using Internet Explorer as their main browser. Google Chrome was labeled as "not accessible" by one of the participants who stated, "I tried it a few times but nope. I tried it a number of times, but theres no talking when you view a website." The participants also explained that Internet Explorer was the browser most commonly taught in schools for the vision-impaired.

Although both participants had an unclear understanding about the concept of a browser extension, when showed the features of the extension, there was an instantaneous positive reaction. Both participants were able to comprehend and enjoy the content shown to them through the different prototypes of the extension, even laughing at the jokes within the video as the description was read aloud by the synthetic voice. As both users previously expressed an great liking of Youtube and an interest in accessing more content related to their interests, they were encouraged with the potential of the extension.

The female interviewee stated, "This works well for the format and I think it's useful. But I don't like that you have to push enter again, because we won't just sit in front of the computer. We like to walk away, like when we are getting ready to go somewhere. Just like with the television. Of course, it depends on what it is." Her partner agreed that although they found the delivery of annotations useful, they did not want to have to constantly pause and unpause the video themselves. All three prototypes of the delivery system had the downfall of requiring an unpause by the user after each annotation. He continued to explain that with current audio description on movies, the description was fitted around the dialogue for a constant flow, although he stated that there were sometimes problems in which the description interrupted the movie. Both participants agreed that for the brief form of Youtube videos, the same style of audio description used in movies might pose even more problems. However, they still insisted that "it would be cool if there was just a button, not pausing, just a button to turn it on and just play."

When asked whether they preferred a live or synthesized voice, both participants agreed that live was preferable. However, the male participant had not noticed that the voice used in the prototypes was a synthetic voice. When told, he expressed satisfaction with the current level of synthetic voice. The other participant also agreed the current level of synthetic voice was sufficient, but ex-

pressed a desire to be able to change the voice, stating, "They should give the user the option of sound type with different features and different voices." She preferred a British male voice while her partner usually preferred a female voice, like the female Siri who he described as having "better pronunciation".

Both participants chose the first and third prototypes as being more effective than the second prototype. The female participant stated "there's more work and guessing with the second one." The first prototype was most strongly preferred due to it's automatic delivery of annotations without requiring the user to press enter to pause the video. However, the participants found the concept of the third prototype to be interesting, because of the level of choice it offered — although that choice was also found to be troublesome, since it would require their immediate attention and proximity to the computer at all times. Depending on the content, the third prototype was said to offer some benefits over the third, but overall the first prototype was foreseen to be the most easy to integrate into their Youtube viewing habits.

4.4.4 Participant 5

The fifth participant in the evaluation study was a male in his early twenties from Japan. Although he did not share the cause of his vision-impairment, he has been vision-impaired since birth, legally blind in both eyes. He was the youngest participant in the study. He was interviewed on July 22, 2014 at a vocational development center in Tokyo, Japan where the researcher volunteers as an English teacher. Other than a keen interest in learning English, the participant also enjoys keeping up with Japanese politics and traveling domestically in Japan. Unlike all previous interviews, the interview had to be cut short due to schedule of the participant.

As a dedicated fan of English-language punk rock bands Sum 41 and Simple Plan, the participant often uses Youtube to listen to music. When asked about what type of videos he likes on Youtube, the participant responded, "Because I am a blind person, I cannot watch a picture. However, I watch music and the animation of the politician only by a sound in Youtube." When shown the Popscriptive prototype and asked whether he would like a system that provided annotations in a similar way to how audio descriptions are provided in movie theatres, the

participant stated, "I intend to continue using Youtube for an animation with or without explanation." In particular, the participant seemed doubtful of whether the system would be worthwhile to use with animation, saying specifically that it seemed "unnecessary" as he believed animation was too "fast-moving" and the "explanation does not catch up with it." Although he did not demonstrate a strong interest in the the prototype concept, the participant did express some dismay at the lack of audio described movies and television dramas available. However, he rejected the idea of user-generated description due to the importance of objectivity. The participant stated, "The commentary must be neutral. This is because you must not classify feelings into the commentary.", but he did not state any objections to the synthesized style and quality of the annotations.

4.5. Secondary User Study

As Popscriptive requires a secondary user group of annotation providers, in addition to the target vision-impaired users who interact with the extension itself, it was crucial to evaluate the system from the perspective of this secondary group. As described in Chapter 3, members of this group will be sighted users of Youtube who input annotation data into the database by sending tweets of descriptive information along with the video ID, start time, and end time. These non-professional annotators will contribute descriptions on a volunteer basis, so it was necessary to investigate their motivation to complete the process.

Nine sighted participants were asked to select a Youtube video of their choice and annotate as much as they felt like doing. An in-depth interview was then conducted about their feelings about the required actions and their willingness to create annotations in the future. Ages and gender of participants included one female in her teens, three females in their twenties, one female in her fifties, three males in their twenties, and one male in his fifties. Although the nationality of participants differed, including the United States, Japan, and Canada, all participants spoke English at a high level of fluency and were asked to make the annotations in English.

Participants were asked to use their own Twitter accounts to tweet, but in the case that they were unwilling to use their personal account, a sample account was

provided for them. Out of the nine users, eight users were able to successfully tweet from their own account while one user tweeted from the sample account. Only one tweet failed at first due to the lack of a comma in the format.

Although participants were willing to create the first annotation, they were not willing to continue creating annotations without any request from the researcher. All participants expressed an interest in the system and praised its potential, but did not foresee themselves actually volunteering their time and effort as annotators. A male participant in his twenties stated, "Its a nice thing. But Im not sure if I would do it on my own volition." A female user in her twenties shared the sentiment, but included that, "I would be motivated if I was assigned small bits. But not like a whole video." When asked about their interest in being part of a volunteer group that would distribute small tasks for annotation, users responded positively but did not necessarily show a deep interest in pursuing the activity in their personal lives.

As for the process of inputting the video ID, start and stop times, and description, participants were able to successfully create a properly formatted tweet after only one demonstration. However, participants were not willing to perform the required operations on their iPhone due to the difficulty of copying and pasting the video ID from the Youtube URL.

4.6. Results

Interviews with the participants were transcribed and evaluated as qualitative data. Data was separated into three main categories in accordance with the goals of the study. In particular, data supporting the second goal of technical quality and user interaction was analyzed by adapting the qualitative data logging template of the SEM-CPU methodology. The adapted data log consisted of five columns, including "Number, "Prototype", "Vision-impairment", "Problems faced", "Classification", and "Design Recommendation" [25]. (See Appendix A.) This data was also used in the third category to realize future design considerations in addition to direct comments made by the participants.

Effectiveness of System for Target User

Four out of five participants responded affirmatively to the basic question of whether they found the tool to be helpful to their content consumption and whether they would want to use the tool in their daily lives when watching. The fifth respondent was content with his current Youtube routine and at the moment was not interested in using the Popscriptive extension.

The overall positive reception by vision-impaired participants in the user study shows that there is a demand for a tool that can provide a deeper engagement experience for the vision-impaired interacting with web video content. All test participants were interested in Youtube and expressed a desire to watch more videos or have annotation available for a certain type of video category they had an existing interest in.

Technical Quality and Interaction Experience for Target User

Response was very positive to the basic form of the annotations (brief descriptions of the main visual cues in the scene) and the synthetic text-to-speech audio presentation. The amateur level quality of the descriptions was also approved by most participants in the study. The acceptance of the basic structure of annotations established the proposed Popscriptive format as a successful method for conveying description information for web video.

One of the most obvious issues that was brought up through the interviews was the low usage of Google Chrome by the vision-impaired participants. In fact, no single participant used Google Chrome due to accessibility issues of the browser itself. Compounding the issue was the low usage of extensions or add-ons. There appeared to a general disinterest in extensions and add-ons among the participants interviewed. Although the Chrome store offers an accessibility section, there may be problems related to how accessible the store is or simply in how the knowledge about add-ons and extensions for the vision-impaired is distributed. Extensions and add-ons are still a relatively new feature of browsers, so it is very likely that their popularity could grow with the vision-impaired population with more exposure.

Participants with a higher level of vision-impairment seemed to show a stronger preference for the automatic delivery system of prototype one, which pauses the

video for the user automatically without waiting for their decision. The same participants also insisted on automating the resume function, although the automatic resume is a feature requested by almost all participants who found it troublesome to make an additional separate action to resume the video. Although the initial pausing not being automated may give the user an option to decide whether the annotation description is necessary for himself or herself personally, the nonautomated resume does not currently hold any benefit of choice.

Participants who retained some level of vision, even if in just one of their eyes, showed a stronger preference for the third prototype. The participants could often make out the shapes and forms in the video but had difficultly seeing the detail. They were able to determine when a detailed visual was most likely important and use the third prototype's on-demand annotation feature.

In general, the need for options and choice was made very clear by all participants in the study. However, all participants found the second prototype with on-demand annotations and no alerts to be too much guesswork and not effective for watching video. Combining the opinions of the five participants, a new prototype would feature both the automated delivery and the alert-style delivery as two options for delivery methods. When a user walks away from the computer or is busy doing another activity, the first delivery method was shown to be very important. However, if the user's full attention is on watching the video, then the third prototype is less intrusive and puts the control in the user's hand.

Results of Secondary User Study

The secondary user study focused on the annotation creators, rather than the consumers. The annotation creators would most likely not need to come in contact with Popscriptive, so the study focuses primarily on the feasibility of volunteers creating annotative tweets while watching a Youtube video The objective of the study was to understand willingness or motivation to provide such tweets, along with substantial barriers.

The evaluation of participants in the secondary user group showed that although participants had positive feelings about the idea of providing annotative descriptions for the vision-impaired, they did not feel motivated enough to provide them on their own volition. If small pieces of work were distributed to them thus

they would only have to annotate a few seconds of video in order to contribute, their willingness to join the effort was increased significantly — but there was still the issue of whether the participants would actually join such a distribution system or volunteer community.

Reactions to the Twitter-based input system were mixed, although participants could immediately learn and use the required format including the video ID, start time, stop time, and annotation with required comma separation. Participants did find the need to copy and paste the video ID to be inconvenient, especially in the case of tweeting from a smartphone or mobile device. When asked whether they would prefer to tweet or use a separate input system, most users responded indifferently to either option but were not particularly positive about using Twitter.

Although the generally affirmative evaluations by both user groups demonstrated proof-of-concept for Popscriptive as an effective tool, many changes still need to be made based on the user study data. In particular, due to the importance of the secondary user group for providing annotations, a more practical or compelling input system for volunteers needs to be researched. Future work for Popscriptive will be discussed in further detail in Chapter 5.

Chapter 5

Conclusion

With the rapid evolution of technology, online culture will most likely continue its visual trend. Consequently, culture as a whole will follow in this visual nature. Social video will increasingly become a significant medium for culture creation, as we can already observe today in the viral success of videos, such as PSY's Gangnam Style, which weave their way into the general consciousness and become an article of cultural knowledge. It is necessary for all people, including vision-impaired users, to have equal access to culture – and therefore, they must have equal access to content. As Brian Charlson, the Chair of the ACB's Information Access Committee stated in a Senate Committee Hearing Committee Hearing entitled *The ADA and Entertainment Technologies: Improving Accessibility from the Movie Screen to Your Mobile Device*, "description provides keys to our culture, allows users to be more engaged and engaging of others with the shared information[1]."

This research proposes Popscriptive as one possible method of providing better accessibility for the vision-impaired to Youtube video content. Popscriptive provides a system of annotation delivery that enables vision-impaired users of Youtube to have a more meaningful engagement experience. Crowdsourced descriptive annotations provide an alternative to visual information, while seamlessly integrating with the video content itself through a browser extension. The vision-impaired user can navigate and interact with Youtube in the same manner, without being forced to migrate to another specialized assistive site resulting in isolation of vision-impaired users from the existing community.

5.1. Discussion of Research Results

According to the findings from the user study, Popscriptive was established as an effective system for providing descriptive annotations to the vision-impaired user. The overall positive reception by vision-impaired participants in the user study shows that Popscriptive has very high potential to be a tool used for daily consumption of web video content by vision-impaired Youtube users. Three prototypes featuring different delivery models were shown to five visually-impaired participants. In-depth interviews were conducted to evaluate the effectiveness of each model based on a set criteria. Participants gave affirmative but mixed results for the different prototypes. However, overall, the analysis of user responses to open-ended questions and free comments made during the concurrent think-aloud demonstrates the effectiveness of the system design, along with the technical quality and usability of the annotations.

It was also shown that casual volunteers have the potential to be an effective source of annotation for web video content. Although participants initially stated their usual preference for recorded voice, they were very open to the idea of synthesized voice during the study and expressed satisfaction with the audio quality of speech-to-text of the browser extension prototypes. For the format of Youtube videos, the user satisfaction demonstrates that synthesized speech-to-text is an acceptable method for providing description to the vision-impaired user and therefore can be used to provide description by amateur annotators with low barriers. Amateur annotators will not be required to record their own voice or have any other editing or technical skill in order to create annotations. To annotate, volunteers will only need to type a line of text. This will also allow Popscriptive's input model to be more potentially scalable to the volume of Youtube content.

A secondary user study on amateur annotators using the Popscriptive system showed that participants are relatively willing to do very small annotation tasks, although they are not willing to commit to a larger sized project. There was an overall positive attitude towards the concept of volunteers providing descriptive annotations of Youtube videos for the vision-impaired, but no concrete motivation was found among the participants. As Popscriptive relies on the input of volunteers for the annotation database, the ambivalent attitude found in the user study demonstrates the need for further research regarding motivation for volun-

teer annotators. Possibilities for this research will be discussed in the following section, along with a general view for the future of the Popscriptive system as a whole.

5.2. Future for Popscriptive

One of the fundamental goals of this research project was to raise awareness about the need for accessibility to video content for vision-impaired users. In particular, a main objective was to spark a discussion for better accessibility and usability for vision-impaired users of Youtube. Ideally this would be a native Youtube tool, rather than an outside extension like Popscriptive.

However, until a collaboration with Youtube can realized or Youtube independently integrates a similar tool within their service, Popscriptive will continue to exist in an extension form. As the user study revealed Google Chrome to be an unpopular browser amongst vision-impaired participants, Popscriptive will be adapted to be released as a Firefox Add-on. Popscriptive will continue to exist in Chrome Extension as the popularity among the general population is still a significant factor that potentially predicts a similar trend among visually-impaired Internet users as Chrome becomes more accessible by screenreader technology. The Extension and Add-on forms are applicable for desktop browsing but cannot be currently used on mobile. In order to provide the same video content experience on the mobile platform, further research may involve creating a smartphone application.

In the user study, participants emphasized the need for options. It is important to continue testing many different methods and details of annotation delivery, as well as implementing customizability options for the user. A future version of Popscriptive may include a way to choose between delivery types, select a specific accent and gender of the synthesized voice, and adjust pace and speed of annotation delivery.

5.2.1 Volunteer Communities

Currently, there exists small communities of people – both vision-impaired and sighted – collaborating for advocation and production of audio description for

online videos. One such community revolves around YouDescribe, a project introduced previously in Chapter 2: Related Works. In the case of YouDescribe, one of the developers of the project actively promotes his project on Twitter, creating hashtags such as #YouDescribe and #VIDesc. A couple particularly interesting hashtag related to promotion of the description cause is the #DescribeAthon14 and #GAAD used to promote a 24-hour marathon of volunteers providing audio description in recognition of GAAD (Global Accessibility Awareness Day).[2] Another hashtag related to the YouDescribe project is #YDRequest, which stands for "YouDescribe Request". This hashtag can be used by vision-impaired Twitter users to request a certain video be described by volunteers. Volunteers interested in describing videos and helping the community can follow this hashtag on Twitter to provide descriptions for requested videos.

A similar Twitter community could be set up to motivate volunteers for the Popscriptive system. The existence of the YouDescribe community proves that there are willing volunteers for providing description for web video, however, the limited reach of current community also demonstrates that the current method of Twitter promotion may not be sufficient for building a large base of volunteers.

While Popscriptive and YouDescribe focus exclusively on Youtube content, The Netflix Accessibility Project is a blog and Facebook group community that focuses specifically on accessibility to video content on the popular online video streaming service, Netflix. The project concentrates on advocacy for accessibility through grass roots movements rather than on creating a tool or technology. Volunteer members of the project write letters to Netflix, as well as letters to production studios and government representatives. The actions of the team are organized and synchronized through updates posted to the blog and Facebook group.

Although both projects have similar goals, the nature of The Netflix Accessibility Project is distinctly diffrentiated from YouDescribe as one recruits volunteers to provide amateur annotations while the other recruits volunteers to solicit professional annotation. However, both volunteer communities through their various social media channels have a relatively limited reach. An effective method for recruiting and motivating volunteers needs to be researched further in order to create a substantial user base that can potentially annotate videos at a rapid

enough speed to keep pace with popular trending videos on Youtube, as well as meeting the requests of individual vision-impaired users.

The secondary user evaluation conducted in correspondence to the prototype Popscriptive system showed the necessity to break up and distribute description jobs into extremely brief parts in order to avoid overextending the volunteer and negatively affecting their motivation to contribute. A system that automatically divides and allocates a volunteer user to a short video section might be a possible solution. Amazon Mechanical Turk is a possible example system to take into reference when designing and prototyping this solution.

In addition, it is necessary to research further about what motivational factors are effective for recruiting and maintaining volunteers as amateur annotators. Non-monetary compensation in the form of online social status rewards should be considered. The same motivating factors, "includ[ing] connecting with peers, achieving a certain level of fame, notoriety, or prestige, and self-expression", that inspire the sharing of video content have a strong plausibility in also motivating the creation of description [50].

Although the Popscriptive prototype utilizes Twitter as a input channel for volunteers to post annotations to the database, Twitter is by no means proven to be the best popular method for input. Although Twitter has a significant user base and can therefore function as a familiar input system for users, participants in the user study were mixed in their willingness to use Twitter as a channel to send the annotation information. The inconvenience of inputting the video ID was also a raised issue – particularly for using a mobile phone to annotate. Possible design solutions for this problem should be researched further, including testing a web form input or native app input system for annotators. In addition, it will be interesting to prototype a native Youtube version of Popscriptive and see how sighted users react to the option of being able to provide descriptive information for vision-impaired users within the platform itself.

It is also important to consider a request system for vision-impaired users. In the current Popscriptive prototype, vision-impaired users interact with the extension and output data of annotations in a relatively passive manner. Annotations are provided as the vision-impaired users browse Youtube, but users cannot interact with the annotations to provide feedback or send requests. As the feedback

and requests will need to be managed by the sighted volunteers, these two systems will need to be integrated in some manner.

5.3. Future for Online Video and Accessibility

Tommy Edison is a Youtube user who has gained wide recognition as one of the world's few blind movie critics. In addition to Youtube, he is an active user of Twitter, Facebook, and Instagram. Rather than letting the visual nature of social media become an obstacle, Tommy instead approaches it from a new perspective, appropriating each medium and channel to turn assumptions around on themselves. In the future, there will be more and more vision-impaired people like Tommy Edison[3] who use social media and the Internet freely.

Presently, popular video services like Youtube and Netflix still ignore their vision-impaired user base. Netflix has even blatantly refused to take responsibility for making content accessible[4]. But the mindset of creating something for the masses and ignoring the individual is rapidly disappearing. While this change is predominately grassroots, as can be witnessed from the W3C Guidelines and other online advocate groups for accessibility and usability, it will also be continuously pushed forward by institutions and the government[5]. Thus it will be critical for content platforms to make their content inclusive.

From a technology perspective, the imminent shift of the Internet from Web 2.0 to the next incarnation will also entail a need for media content to be better labeled, described, categorized, annotated, and made more meaningful for both human and computers. Although current computer vision research revolves predominately around creating a better world for sighted people, it also holds endless promise for the vision-impaired people. One such promise may be automated video annotation.

Vision-impaired people are highly interested in a wide variety of Youtube content but desire more accessibility. Accessibility should not just be defined as the means of entry, but also as a means of comprehension. Accessibility without usability and usability without accessibility are both incomplete concepts. "To the greatest extent possible, people with disabilities would like to use the same applications, the same tools, and the same devices, and access the same content as

their peers without disabilities [29]."

Popscriptive is one tool to provide a deeper engagement experience through more meaningful content for vision-impaired users on Youtube. In order to push the accessibility movement forward, it is important for there to be options and choices for the vision-impaired user, but it is also crucial to avoid fragmenting the data and the volunteer effort. With the long-term goal of Youtube native implementation, Popscriptive future aims include seamless annotation delivery on Firefox, Chrome and mobile browsers, customizable playback features, and an effective system for motivating and managing volunteer annotators. In addition, the research conducted for Popscriptive can be used for developing new methods of annotation delivery and new styles of video for the vision-impaired user.

Notes

1 http://www.help.senate.gov/hearings/hearing/?id=0a89258a-5056-a032-5276-85441c3431e8

2 https://twitter.com/BerkeleyBlink

3 www.youtube.com/user/TommyEdisonXP

4 http://netflixproject.wordpress.com/2014/03/25/netflix-keeps-refusing-accessibility-and-audio description/comment-page-1/comment-301

5 http://washear.org/restoration_act.htm

Acknowledgements

First and foremost, I would like to express my sincere appreciation to Okude-sensei and Inakage-sensei for their ceaseless guidance throughout my time as a student at KMD. Having the chance to participate in both OIKOS and PLAY Projects during these past two years has been an invaluable life experience that I will never forget or take for granted. I would like to say a special thanks to Okude-sensei for his weekly writing workshops without which this thesis would not have been possible. I also need to extend my gratitude to Minamizawa-sensei for his role as my third supervisor. His comments during my midterm and final presentation helped me immeasurably in finding focus for my research.

I need to say thank you to Inakage-sensei (once again!) and Ueki-sensei for their patience and supervision in PLAY Project. Under PLAY, I was able to pursue my own goals while learning from my fellow project members. In particular, I want to say thanks to Gaby for being my co-member in the Social Video subproject. I would be remiss to forget to say my thanks to Shikei (Zihui), Tomo, Pan, Toshi, and, especially, Kazuma for their perpetual help and friendship.

I cannot say thank you enough to all the participants for their time and feedback. I would also like to thank CWAJ for providing me with the opportunity to forge a connection with the VI community. All of you ladies are so sweet and thoughtful. Another organization I cannot miss in my gratitude is Japan MEXT for their generous support, along with the KMD Office and the MEXT Coordinators. I am also profoundly indebted to Dean Tyler Stovall, Director Roseanne Fong, and Professor John Wallace from University of California, Berkeley who all went above and beyond to assist me in preparing for my journey to Japan.

I would like to give a shout out to my September Batch! (Although I am quite fond of the April students too. Yuki, Sanchuu, ex-Kara/AmbiFurn: Moeko,

Kisshan, Makoto, Kohsuke!) I truly believe my happiness and success at KMD has been because of your friendship. Thank you to Ivy and the rest of Team E Ne (Thuy, Keita, & Kiron) for being there with me from the very beginning. (Ivy, I hope to continue being your minion forever!) I am so happy that I have made friends that will last a lifetime here. I could never forget my nighttime conbini buddies, Carlos and Mariam — with a tiny extra love for Mariam because she deserves a tiny extra everything. I am really grateful for everyone but I just want to give a quick mention to Kailin, Flash, Pim, Tina, and Niya!

Outside of school, I would like to thank Goodpatch and Tsuchiya-san for taking a chance on me. I think now would be a good time for my hat tip to Boris, as well. I am so happy that you have become a big part of my life, even if sometimes I complain.

Thanks to Tracey and Krizette, my favorite Rosevillians! I know you have been very content with the distance between us. :)

Last, but far from least, I would like to thank my parents who I love with all my heart and who I know love me even more than that. I am so blessed to have such a supportive and caring "parental unit". The best thing that ever happened to me was being your daughter. Thank you for the nagging and the daily attempts at interesting news updates. I love you.

Less than a week before midterm presentations, my father was diagnosed with lymphoma cancer. If it wasn't for the support of everyone at school and for the extra encouragement from both my mother and father, my life would have fallen apart the instant I heard the news. Instead, I was able to successfully finish writing this thesis and enjoy everyday at KMD to the fullest.

References

[1] Abeele, M. V., De Cock, R., and Roe, K. Blind faith in the web? internet use and empowerment among visually and hearing impaired adults: a qualitative study of benefits and barriers. 129–151.

[2] Ashok, V., Borodin, Y., Stoyanchev, S., Puzis, Y., and Ramakrishnan, I. V. Wizard-of-oz evaluation of speech-driven web browsing interface for people with vision impairments. In *Proceedings of the 11th Web for All Conference*, ACM (2014), 12.

[3] Azenkot, S., Prasain, S., Borning, A., Fortuna, E., Ladner, R. E., and Wobbrock, J. O. Enhancing independence and safety for blind and deaf-blind public transit riders. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM (2011), 323–324.

[4] Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., and Miller, R. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM (2010), 333–342.

[5] Brabham, D. C. *Crowdsourcing*. MIT Press, 2013.

[6] Branje, C. J., and Fels, D. I. Livedescribe: Can amateur describers create high-quality audio description? 154–165.

[7] Burgess, J., and Green, J. *YouTube: Online video and participatory culture*. John Wiley Sons, 2013.

[8] Burton, Michele A., E. B. R. B. C. N. J. P. B., and Hurst, A. Crowdsourcing subjective fashion advice using vizwiz: challenges and opportunities. In

Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, ACM (2012), 135–142.

[9] Buxton, B., and Marquardt, N. *Sketching User Experiences*. Elsevier, 2012.

[10] Dinesh, T., Uskudarli, S., and Choppella, S. V. Re-narration as a basis for accessibility and inclusion on the world wide web. In *Proceedings of the 9th International Cross- Disciplinary Conference on Web Accessibility, Lyon, France* (2012).

[11] Doan, A., Ramakrishnan, R., and Halevy, A. Y. Crowdsourcing systems on the world-wide web. 86–96.

[12] Encelle, B., Ollagnier-Beldame, M., Pouchot, S., and Prie, Y. Annotation-based video enrichment for blind people: a pilot study on the use of earcons and speech synthesis. In *Proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ACM (2011).

[13] Fels, D., Udo, J., J.E., D., and Diamond, J. A first person narrative approach to video description for animated comedy. 295–305.

[14] Fels, D. I., Udo, J. P., Ting, P., Diamond, J. E., and Diamond, J. I. Odd job jack described: a universal design approach to described video. 73–81.

[15] Ferreira, S. B. L., Nunes, R. R., and Silva da Silveira, D. Aligning usability requirements with the accessibility guidelines focusing on the visually-impaired. 263–273.

[16] Gagnon, L., Chapdelaine, C., Byrns, D., Foucher, S. F., Heritier, M., and Gupta, V. A computer-vision-assisted system for videodescription scripting. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference*, IEEE (2010), 41–48.

[17] Hansen, A. D. Narrating the game: Achieving and coordinating partisanship in real time. 269–302.

[18] Hills, P., and Argyle, M. Positive moods derived from leisure and their relationship to happiness and personality. 523–535.

[19] Hocking, C. Function or feelings: factors in abandonment of assistive devices. 3–11.

[20] Howell, J. T. *Hard living on Clay Street: Portraits of blue collar families*. Anchor Press, 1973.

[21] Kelly, B., Sloan, D., Phipps, L., Petrie, H., and Hamilton, F. Forcing standardization or accommodating diversity?: a framework for applying the wcag in the real world. In *Proceedings of W4A05*, ACM (2005), 46–54.

[22] Kobayashi, M. "unifying video captions and. text-based audio descriptions". In *CSUN 2011, IBM Research Tokyo* (2011).

[23] Lai, P. P. Application of content adaptation in web accessibility for the blind. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, ACM (2011).

[24] Lazar, J., Dudley-Sponaugle, A., and Greenidge, K.-D. Improving web accessibility: a study of webmaster perceptions. 269–288.

[25] Lee, Y. S., Hong, S. W., Smith-Jackson, T. L., Nussbaum, M. A., and Tomioka, K. Systematic evaluation methodology for cell phone user interfaces. 304–325.

[26] Loewen, G., and Tomassetti, V. Fostering independence through refreshable braille. In *Presentation at the Developing Skills for the New Economy: International Conference on career/technical and Vocational Education and Training, Manitoba* (March 2006).

[27] Lopes, R., Gomes, D., and Carrico, L. Web not for all: A large scale study of web accessibility. In *Proceedings of W4A10*, ACM (2010), 10.

[28] Mander, J. Gwi social q2 2014: the latest social networking trends. GlobalWebIndex, May 2014. http://blog.globalwebindex.net/gwi-social-q2-2014.

[29] Manduchi, R., and Kurniawan, S. *Assistive technology for blindness and low vision*. CRC Press, 2012.

[30] McManus, R. The rise of beautiful apps. ReadWrite, April 2012. http://readwrite.com/2012/04/30/the-rise-of-beautiful-apps.

[31] Mereu, S. W., and Kazman, R. Audio enhanced 3d interfaces for visually impaired users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (1996), 72–28.

[32] Newell, A. F., Gregor, P., Morgan, M., Pullin, G., and Macaulay, C. User-sensitive inclusive design. 235–243.

[33] Packer, J., and Kirchner, C. Who's watching?: A profile of the blind and visually impaired audience for television and video.

[34] Phillips, B., and Zhao, H. Predictors of assistive technology abandonment. 36–45.

[35] Poeter, D. Facebook introduces revamped visual news feed. PCMagazine, March 2013. http://www.pcmag.com/article2/0,2817,2416361,00.asp.

[36] Power, C., Freire, A., Petrie, H., and Swallow, D. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012).

[37] Prasain, S. Stopfinder: improving the experience of blind public transit riders with crowdsourcing. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM (2011), 323–324.

[38] Rajapakse, R., Dias, M., Weerasekara, K., Dharmaratne, A., and Wimalaratne, P. Audio user interface for visually impaired computer users: in a two dimensional audio environment. In *Proceedings of World Academy of Science, Engineering and Technology, no. 65. World Academy of Science, Engineering and Technology* (2012).

[39] Rotman, D., and Preece, J. The 'wetube' in youtubecreating an online community through video sharing. 317–333.

[40] Schmeidler, E., and Kirchner, C. Adding audio description: Does it make a difference?

[41] Shontell, A. Vine is the fastest-growing app of 2013. BusinessInsider, October 2013. http://www.businessinsider.com/vine-is-the-fastest-growing-app-of-2013-2013-10.

[42] Sloan, D., Gregor, P., Rowan, M., and Booth, P. Accessible accessibility. In *Proceedings on the 2000 conference on Universal Usability*, ACM (2000), 96–101.

[43] Snyder, J. Audio description: The visual made verbal. 935–939.

[44] Soderstrom, S. "digital differentiation in young peoples internet useeliminating or reproducing disability stereotypes.". 190–204.

[45] Spencer, L. *Quality in Qualitative Evaluation: A framework for assessing research evidence*. Government Chief Social Researcher's Office, Cabinet Office, 2003.

[46] Udo, J. P., Acevedo, B., and Fels, D. I. Horatio audio-describes shakespeares hamlet blind and low-vision theatre-goers evaluate an unconventional audio description strategy. 139–156.

[47] Udo, J. P., and Fels, D. I. Re-fashioning fashion: an exploratory study of a live audio-described fashion show. 63–75.

[48] Unknown, U. Statistical snapshots from the american foundation for the blind. American Foundation for the Blind.

[49] Unknown, U. Screen reader user survey 5 results. WebAim, February 2014. http://webaim.org/projects/screenreadersurvey5/.

[50] Vickery, G., and Wunsch-Vincent, S. *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD), 2007.

[51] Vondrick, C., Patterson, D. P., and Ramanan, D. Efficiently scaling up crowdsourced video annotation. 184–204.

[52] Winograd, T. From programming environments to environments for designing. 69.

[53] Zillmann, D., and Vorderer, P. *Media entertainment: The psychology of its appeal*. Routledge, 2000.

Appendix

A. Qualitative Data Log

Table 4.2: Technical quality and user interaction issues from user study.

Number	Prototype	Vision-Impairment	Problems faced	Classification	Design recommendation
1	1,2,3	ALL	Do not know how to use Chrome Extension.	Interaction	Consider using Firefox Add-On or alternative to provide same tool functions.
2	1,2,3	ALL	The need to pause the video after annotation to continue playing is annoying.	Mental Model	Make resume automatic after annotation is finished being read by text-to-speech.
3	2	ALL	Too many failures. (Keep receiving "Annotation not available" message.)	Interaction	Develop an algorithm or alternative method to limit failure or make failure less disturbing to the user.
4	1,2,3	Low	Buttons too small on player. Cannot see.	Interaction	Make buttons larger/resizable or eliminate button interaction.
5	1	Med	Sound of alert tone chiming is annoying.	Mental Model	Create a more subtle tone or give the option to turn tones on/off, customize sound.
6	1	Low	Too hard to hear sound of alert tone.	Interaction	Higher volume option for tones.
7	1,2,3	ALL	Timing of annotations is sometimes off.	Mental Model	Develop a more accurate way to apply time stamps to annotations, perhaps detecting audio of original video. Need quality control for annotation time stamps.
8	1,2,3	ALL	Video moves too quick, cannot pause annotation in time.	Interaction	There may need to be a delay for human latency in delivering the annotation.