

**COLLABORATIVE ANNOTATION, ANALYSIS, AND PRESENTATION  
INTERFACES FOR DIGITAL VIDEO**

A Dissertation  
Presented to  
The Academic Faculty

by

Nicholas A. Diakopoulos

In Partial Fulfillment  
of the Requirements for the Degree  
Ph.D. in Computer Science in the  
School of Interactive Computing, College of Computing

Georgia Institute of Technology  
August, 2009

**COPYRIGHT © NICHOLAS A. DIAKOPOULOS**

**COLLABORATIVE ANNOTATION, ANALYSIS, AND PRESENTATION**  
**INTERFACES FOR DIGITAL VIDEO**

Approved by:

Dr. Irfan Essa, Advisor  
School of Interactive Computing  
*Georgia Institute of Technology*

Dr. Cliff Lampe  
Department of Telecommunication,  
Information Studies and Media  
*Michigan State University*

Dr. Gregory Abowd  
School of Interactive Computing  
*Georgia Institute of Technology*

Dr. John Stasko  
School of Interactive Computing  
*Georgia Institute of Technology*

Dr. Jay Bolter  
School of Literature, Communication, and  
Culture  
*Georgia Institute of Technology*

Date Approved: May 18, 2009

## **ACKNOWLEDGEMENTS**

There are many people that have contributed to my success over the years, not just on this piece of work but also in my broader intellectual and personal development. It's been a journey and I can say that I'm certainly not the same person I was when I first started this endeavor.

First and foremost I want to acknowledge and thank Kathryn both for keeping me whole and for her patience and help in the diligent pursuit of my goals. I also owe a huge debt to my media savvy family (Mom, Dad, Chris, and Dimi) for forging me the way I am and for their individual expertise in all different aspects of media. They have informed me in ways that I could not otherwise be. My friends have pushed me the whole way, making me stronger with each intellectual conversation and each new perspective.

I'm impelled to thank Irfan in particular here especially for his effort in creating opportunities that have allowed me to grow and mature as an individual and as a researcher. And to the rest of my committee for providing their feedback, expertise, and insights as I found my way.

Early on I was molded by the senior students in CPL: Gabe, Vivek, and Drew, all of whom were generous with their time. Even before that I have to acknowledge Mike Sullivan, Mike Natalizia, Tony Lima, and Chris Natalie at Astro-Med, who created my first opportunity for doing HCI. Thanks also to my other early mentors, Peter Stephenson and Stephan Volmer, who taught me the value of perseverance.

And finally, a huge thanks to the coffee shops of Atlanta, who have kept me awake though the long slog: Inman Perk, SFCR and all the rest!

## TABLE OF CONTENTS

|   | Page |
|---|------|
| ACKNOWLEDGEMENTS  | iv   |
| LIST OF TABLES  | vii  |
| LIST OF FIGURES   | viii |
| NOMENCLATURE AND ABBREVIATIONS  | ix   |
| SUMMARY   | x    |
| <u>CHAPTER</u>  |      |
| 1 Introduction  | 1    |
| Goals and Approach  | 3    |
| Thesis Statement and Contributions  | 6    |
| Overview of Dissertation  | 9    |
| 2 Multimedia Annotation Methods and Interfaces: Related Research              | 12   |
| Fully Automated Annotation  | 12   |
| Manual and Semi-Automated Annotation Interfaces                               | 14   |
| Perceived Costs of Annotation   | 23   |
| Summary   | 27   |
| 3 Annotation of Audio-Visual Media with Time-Stamped Transcripts Using a Game | 29   |
| Audio Puzzler Game Design   | 30   |
| Transcription Algorithm   | 33   |
| Laboratory Evaluation   | 34   |
| Online Deployment and Evaluation  | 41   |
| Discussion  | 51   |

|   |  |     |
|---|--|-----|
| 4 | An Annotation Model for Making Sense of Information Quality in Online Video                | 56  |
|   | Information Quality and Its Constituents   | 57  |
|   | Practical Design Goals   | 62  |
|   | Information Quality Operationalization   | 64  |
| 5 | Videolyzer: Annotation of Information Quality in Online Video for Bloggers and Journalists | 68  |
|   | Related Work   | 69  |
|   | Videolyzer System  | 72  |
|   | Videolyzer Interface   | 76  |
|   | User Study   | 82  |
|   | Discussion   | 92  |
| 6 | Modulating Video Credibility Through Annotation Syndication and Simplified Presentation    | 95  |
|   | Related Work   | 95  |
|   | Visualization Design   | 97  |
|   | Experiment   | 99  |
|   | Discussion   | 105 |
| 7 | Conclusions and Future Work  | 107 |
|   | Limitations  | 109 |
|   | Broader Implications and Impact  | 113 |
|   | REFERENCES   | 119 |
|   | VITA   | 129 |

## LIST OF TABLES

|   | Page |
|---|------|
| Table 1.1. A summary of research questions, research activity, and contributions..            | 11   |
| Table 3.2. Summary Statistics for Game Play Ratings (7 points scale).....                     | 37   |
| Table 3.3. Summary Statistics for Transcription Accuracy. ....                                | 39   |
| Table 3.4. Audio Puzzles Used for Online Experiment .....                                     | 46   |
| Table 4.5. Tags for Describing Information Quality and Bias.....                              | 67   |
| Table 5.6. Annotation types and associated actions depending on annotation<br>ownership. .... | 73   |

## LIST OF FIGURES

|  | Page |
|--|------|
| Figure 3.1. Overview of the Audio Puzzler Interface.....   | 31   |
| Figure 3.2. Merging Two Audio Chunks. Optimal overlap point shown in solid red)  | 34   |
| Figure 3.3. Shows the true and measured means for the number of words recorded in a real set of exemplars. The light shaded range shows the area where the exemplars were retained for further processing.....   | 43   |
| Figure 3.4. A buoy (red) is selected from a set of five transcripts by minimizing the distance between the buoy and the other transcripts based on the WER metric. ....  | 44   |
| Figure 3.5. From a corpus of seven transcripts, (a) the buoy is chosen and words deleted which don't have corroboration from the other transcripts. In (b) words from several other transcripts in the corpus are inserted in the buoy to improve its accuracy. .... | 46   |
| Figure 3.6. Aggregation algorithm results showing WER comparisons to ground truth for each case.....   | 47   |
| Figure 3.7. Aggregation Subset Results.....  | 51   |
| Figure 4.1. Key Information Quality Facets.....  | 59   |
| Figure 4.2. Relationships Between Types of Annotations.....  | 66   |
| Figure 5.1. The tagging panel UI with quality, bias, and collaboration tags.....   | 74   |
| Figure 5.2. Overview of the Videolyzer interface showing (clockwise from upper left) the tool bar, filter options, transcript, timeline, annotation view, and video playback area.....   | 77   |
| Figure 5.3. Color coded, automatically detected markers shown on the transcript view.....  | 78   |
| Figure 5.4. Stacked timeline view showing levels for navigation, stripe image, automatically extracted markers, and anchors. ....  | 79   |
| Figure 5.5. The Annotation View showing a claim that has lost support based on the series of annotations responding to it. ....  | 80   |
| Figure 5.6. Annotation Type Usage and Distribution.....  | 88   |
| Figure 5.7. User Experience Ratings.....   | 92   |
| Figure 6.1. Experimental video player showing annotation polarity graph and detail panes. ....   | 98   |
| Figure 6.2. Mean overall credibility ratings for each video and across engagement levels.....  | 103  |

## NOMENCLATURE AND ABBREVIATIONS

|         |  |
|---------|--|
| AMT     | Amazon Mechanical Turk   |
| ASR     | Automatic Speech Recognition   |
| CAPTCHA | Completely Automated Public Turing test to tell Computers and Humans Apart |
| DTC     | Direct To Consumer   |
| ESL     | English as a Second Language   |
| FDA     | Food and Drug Administration   |
| GT      | Ground Truth   |
| IBIS    | Issue Based Information System   |
| OCR     | Optical Character Recognition  |
| POS     | Part Of Speech   |
| PR      | Public Relations   |
| TTS     | Text To Speech   |
| UGC     | User Generated Content   |
| VNR     | Video News Release   |
| WER     | Word Error Rate  |

## SUMMARY

Information quality speaks to a degree of excellence in communicating knowledge or intelligence and encompasses aspects of validity, accuracy, reliability, bias, transparency, and comprehensiveness among others. Professional news, public relations, and user generated content alike all have their own subtly different information quality issues. With so much recent growth in online video, it is also apparent that more and more consumers will be getting their information from online videos and that understanding the quality of video information becomes paramount for a consumer wanting to make decisions based on it.

This dissertation explores the design and evaluation of collaborative video annotation and presentation interfaces as motivated by the desire for better information quality in online media. We designed, built, and evaluated three systems: Videolyzer, Audio Puzzler, and Videolyzer CE which contribute both to interface methods for video annotation and to mechanisms for enhancing objective metadata such as transcripts as well as subjective notions of information quality of the video itself.

Videolyzer is a semi-structured manual analysis system for a video, its transcript, and its annotations which was designed to aid bloggers and journalists collect, aggregate, and share analyses of the information quality of a video. Its interface design and evaluation explored many questions of general interest to video annotation including: *granularity*, *transcript integration*, *argumentation systems*, and *automation*. The construction of Videolyzer also entailed adequately defining information quality and operationalizing it as a set of annotations available

to jumpstart people's analyses. We evaluated Videolyzer in a laboratory study and found that users' awareness and understanding of comprehensiveness, multiple perspectives, context, and quality of the video were enhanced.

One component of the evaluation of the Videolyzer interface was the effect that the inclusion of a time synchronized transcript would have on the user experience. This in turn motivated the need for a high quality time-stamped transcript of the video that would allow for interactions with the transcript to be mirrored on the video timeline. Because the accuracy of automatically produced transcriptions of video is generally poor under real-world conditions we developed Audio Puzzler, a game which as a by-product of play produces time-stamped transcripts of videos. We show that Audio Puzzler is an engaging game and further demonstrate a high accuracy of transcript metadata by leveraging an aggregation algorithm to merge many people's independently produced output.

Finally, once we had built an integrated system for analyzing videos with transcripts, we asked the question whether and to what degree the credibility of video information could be modulated for end consumers by these collected annotations. Our goal was to syndicate the knowledge collected using Videolyzer to a class of users that was less engaged, but that would still find benefit in having the additional annotation information. To do this we built and evaluated the credibility impact of simplified visualizations which showed annotation activity, polarity, and sources packaged into an online video player. Our evaluation showed that these graphics could influence people's perceptions of the credibility of a video with stronger effects exhibited for more engaged people.

## CHAPTER 1

### INTRODUCTION

On October 3<sup>rd</sup> 2008, the Food and Drug Administration (FDA) sent a warning letter to Bayer Pharmaceuticals. The FDA claimed that Bayer had run broadcast advertisements for one of its birth control formulations, Yaz, in violation of U.S. laws protecting consumers from misleading or potentially harmful information in DTC (Direct To Consumer) pharmaceutical ads. The problem was that the ads encouraged the use of the drug in circumstances other than those that it was approved for and that they over-promised the benefits and minimized the risks associated with the drug [39]. As punishment? Bayer had to launch a \$20 million dollar ad campaign in 2009 to counter the damage of the misleading ads and correct people's perceptions of the drug [104].

Consider another example in the domain of health-related video, but this time with *user generated content* (UGC), video of the type that ordinary non-experts author and post to sites like YouTube. A content analysis study conducted in 2007 at the University of Toronto [54] found that of 153 videos referring to vaccination or immunization on YouTube, 32% were negative and 20% more were ambiguous with reference to Canadian governmental health guidelines on the topic. Exacerbating the issue was that the researchers found that the negative videos had a higher mean user rating as compared to the positive videos. This is a frightening proposition considering many may believe this contradictory information seen online.

The media issue exposed through these vignettes is one of *information quality*, the degree of excellence in communicating knowledge or intelligence. And

the problem isn't just relegated to marketers misleading or proles producing on YouTube. Studies of newspapers have found information quality problems downright prevalent; 61% of local news and features stories across a sample of fourteen newspapers were found to have some form of error including misquotes, or inaccuracies with names, addresses, locations, or times [75]. And then there's the problem with VNRs (Video News Releases), potentially biased PR (public relations) video segments that broadcast news stations can pick up digitally and drop onto the evening news with minimal editing effort [73, 77].

Information quality encompasses aspects of validity, accuracy, reliability, bias, transparency, and comprehensiveness. Professional news, public relations, and user generated content alike all have their own subtly different sources of information quality issues. Subjectivity and bias are inherent to many information domains and can creep into professionally produced media both from internal editorial pressure or from external advertiser pressure, not to mention the intentional manipulations of public relations produced media [38]. For user generated content add to that a potential lack of editorial guidance and concern for reputation building and there are even more opportunities for sloppy mistakes and uncorroborated information being produced and disseminated.

With so much growth in online video in the last few years, it is apparent that more and more consumers will be getting their information from online videos. Topics like health, finance, and politics are already common place; not to mention the plethora of video advertisement and PR that most people are

incidentally exposed to. Understanding the quality of video information becomes paramount for a consumer wanting to make decisions based on it.

There are a number of journalistically inclined web sites that have evolved to start addressing issues of information quality in the media. These include non-partisan examples such as FactCheck, Politifact, Washington Post Fact Checker, and Consumer Reports AdWatch, as well as partisan counter-parts such as Newsbusters and MediaMatters. Oftentimes referred to as media *watchdogs* these sites comb the media and engage in fact checking and re-contextualization of news and other media reports. These analyses, typically presented as long blocks of text, could be thought of as annotations or addenda on the original story. Where this approach toward media watchdogging fails on a user-interface level is for time-based media such as video, where the visual and auditory context is divorced from the re-contextualization provided by the textual analysis. To a large extent, interfaces that support this form of watchdogging specifically tailored to the medium of video are non-existent.

### **Goals and Approach**

The primary goal of this dissertation is to describe the development and study of novel collaborative video annotation interfaces which begin to address the issue of information quality in online video. And while motivated by the overarching information quality context, much of the interface work developed here can be generalized to apply to other annotation contexts with different constraints and scheme of annotation.

*Collaboration* arises as a guiding principle for the development of the annotation systems described here. In traditional journalism practice, information quality is managed rigorously by a reporter collaborating with senior editors. In online knowledge production environments such as those found on Wikipedia, a somewhat different collaborative model based on the *online* and *distributed* collaborative evaluation of information quality is employed [111]. Studies of the quality of entries on Wikipedia have been shown to be similar to those in expertly produce encyclopedias, at least in some content areas [42, 97]. In the Wikipedia model, the process of quality evaluation and assurance is apportioned across many individuals with disparate expertise and interests. Individual knowledge leads people to have different predictions, expectations, and evaluations of information on the web [94], the aggregation of which can enhance media with the context needed for clear understanding.

Furthermore, models developed in media economics [40] suggest that as feedback about the state of information presented increases, the amount of bias occurring in equilibrium decreases, assuming that producers of media have a desire to build a reputation as a high quality source. As a method of providing feedback from an array of different evaluators (such as is done on Wikipedia), collaborative annotation stands as a viable approach toward reducing bias in equilibrium.

Annotation of *video* in particular represents a range of difficult problems at the interface level. For instance, the representational issues associated with a time-varying signal across dual sensory channels (visual and auditory) leads to challenges associated with navigation, selection, and segmentation. Other problems

include how to reduce actual or perceived costs for the user (e.g. by reframing it as a game), how to integrate automation and semi-automation into the interface, and how users can effectively collaborate around the annotation task. It is these interface level problems which form the basis for this dissertation, with the development of the annotation scheme for information quality serving as an important contextual and application constraint on design.

Motivated by information quality concerns in the media and guided by a collaborative annotation approach we have designed, built, and evaluated Videolyzer. Videolyzer's design goal is to aid bloggers and journalists analyze the quality of online political video and then aggregate and share these analyses. It does this by providing a semi-structured manual analysis system for a video, its transcript, and its annotations. The interface of Videolyzer explores many questions of general interest to video annotation including: *granularity*, *transcript integration*, *argumentation systems*, and *automation*. The construction of Videolyzer also entailed the challenge of adequately defining information quality and operationalizing it as a set of annotations available to jumpstart people's analyses.

One of the research questions associated with the evaluation of Videolyzer was to what extent the integration of a textual transcript would effect interaction with and analysis of the video. This in turn motivated the need for a high quality time-stamped transcript of the video that would allow for interactions with the transcript to be mirrored on the video timeline. Since the accuracy of automatically produced transcriptions of video is generally poor under real-world conditions we developed a novel annotation system based on a different approach – a puzzle game

called Audio Puzzler. The game design and evaluation serves as a feasibility proof for the collection of time-stamped transcript annotations using a game approach.

Finally, once we had built an integrated system for analyzing videos with transcripts, we wanted to ask the question whether and to what degree the credibility of video information for end consumers could be modulated by these collected annotations. We felt this was especially important because the high degree of effort and motivation needed to use Videolyzer would preclude other more casual users from reaping the benefits of the somewhat complex argumentation system and presentation. In essence our goal was to syndicate the knowledge collected using Videolyzer to a class of users that was less engaged, but that would still find benefit in having the additional annotation information. To do this we built and evaluated the credibility impact of simplified visualizations which show the semantics of collected annotations (e.g. positive or negative response) packaged into an online video player.

### **Thesis Statement and Contributions**

This dissertation seeks to explore the design and evaluation of collaborative video annotation and presentation systems as motivated by the desire for better information quality. The thesis statement for this research follows:

*Collaborative video annotation interfaces incorporating features related to granularity, transcript integration, argumentation systems, suggestive semi-automation, and ludic engagement are effective methods of video annotation and can enhance both the objective information quality of metadata such as transcripts as well as subjective notions of information quality of the video itself.*

We validated this hypothesis by undertaking the design and construction of Audio Puzzler, Videolyzer, and Videolyzer VCE as well as their subsequent evaluation through user studies. Table 1.1 summarizes the research questions, research activities, and contributions in this dissertation. More extensively stated the contributions of this dissertation are:

- Audio Puzzler represents an existence proof for the efficacy of game-based methods for the collection of highly accurate time-stamped transcript annotations of a video. Although games have been used for the annotation of other media such as images and music [22, 115, 119, 122], this is the first game known to us that accomplishes the transcription of video using a game-based method. Based on both a laboratory evaluation as well as observation of the use of the game online, we contribute insight into the strengths and weaknesses of this method of annotation such as content effects on game play, contextual effects, player knowledge, and scalability. Finally, we contribute an algorithm that can aggregate the transcripts of many different people independently solving an audio puzzle in order to produce a transcript that is 99% accurate.
- In the process of designing Videolyzer we developed a practical information quality model which distills the facets of information quality of most interest to us into a structured set of quality tags and related interactions. In our evaluation of Videolyzer we found this semi-structured approach toward information quality annotation to be usable and expressive for bloggers and journalists. We see this model as a design contribution that could be

implemented by others interested in similar applications but oriented toward different media such as audio or text.

- We also contribute insights relating to the novel features that were implemented and evaluated in a laboratory evaluation of Videolyzer. Our results show that the transcript integration and coupling to the video interactions was found to be a usable and preferable method for navigating and annotating audio information centric video. We also found that users appreciated the granularity of annotation available with the tool as they felt it helped them focus the discussion. Furthermore, the integration of automatically initiated suggestions to aid navigation and annotation were found to have low to moderate user benefit. Finally, we characterize the discussion produced via the annotations in our study and found that in the process of using Videolyzer to do video analysis users' perception of comprehensiveness, perspectives, context, and quality of the video were enhanced.
- We also contribute a presentation design for the annotations collected using Videolyzer so that a simplified set of information quality cues can be presented in line with the video as it is watched. We found that a video player, augmented with graphics showing annotation activity, polarity, and sources could influence people's perceptions of the credibility of the video with greater effects exhibited if they chose to interact with the graphics. We contribute an understanding of the limitations associated with this technique in terms of content and user experience.

## Overview of Dissertation

In the following chapter we will cover related work pertaining to different approaches toward media annotation including automation, reducing actual costs of manual annotation through usability, semi-automation, and collaboration, and reducing perceived costs of manual annotation. The related work reveals areas of potential interest in the design of video annotation systems including what role automation can play in a manual analysis system, how collaboration can be effectively supported, and how games can be designed to reduce perceived costs of annotation.

Chapter 3 details the design and evaluation of Audio Puzzler, a puzzle game which as a by-product of play produces time-stamped transcripts of video. We describe a laboratory study looking at the playability and design limitations of the game as well as an evaluation of the accuracy of the transcripts produced using data collected from online players.

In Chapter 4 we synthesize a range of research across journalism, information science, social science, health information systems, and semiotics to arrive at an operationalization of information quality as a simplified set of tags. Using these tags we also present an interaction model that supports the evaluation of most facets of information quality.

The semi-formal information quality model described in Chapter 4 is implemented in Videolyzer, whose design and evaluation we describe in Chapter 5. We detail the different features implemented in Videolyzer including granularity of annotation, textual transcript integration, argumentation systems, and suggestive

automation as well as the reaction of bloggers and journalists to the usefulness of these features in a laboratory study.

In Chapter 6 we extend the power of Videolyzer by using the rich annotations collected using it to drive more simplified visualizations that are presented in line with the video as it is watched by end consumers. We detail the evaluation of these visualizations and their effect on the credibility of the underlying video information. We conclude the dissertation in Chapter 7 with a discussion of results, contributions, and future work.

**Table 1.1. A summary of research questions, research activity, and contributions.**

|              |   |
|--------------|---|
| <b>RQ 1</b>  | Can a lusrory interface be effectively used to collect accurate transcript and word time-stamp annotations for videos? If so, what are the limitations and benefits of such an interface?   |
|              | <p><b>Research Activity</b><br/>Design and iterative prototyping of Audio Puzzler; Evaluation of Audio Puzzler through user study and accuracy comparison to manually generated ground truth and state of the art automatic methods.</p> <p><b>Contributions</b><br/>Existence proof for the efficacy of game-based methods for the collection of accurate time-stamped transcript annotations of video. Insight into content effects, player knowledge, and scalability as they relate to annotation games. An algorithm for the aggregation of transcripts produced by different people solving the puzzle in order to increase overall accuracy.</p> |
| <b>RQ 2</b>  | How can the many elements of information quality be practically defined, operationalized, and simplified as a set of annotations?   |
|              | <p><b>Research Activity</b><br/>Extensive literature review in information science, journalism, health information, semiotics. Modeling and simplification of information quality into a set of tags for intended user group.</p> <p><b>Contributions</b><br/>The operationalization of information quality was found to be usable and expressive for bloggers and journalists. The model design could be used by other similar argumentation systems.</p>  |
| <b>RQ 3a</b> | Can a usable tool which reifies the model developed in RQ 2 as an argumentation system be used to enabled bloggers and journalists to have a more thorough understanding of the quality of information in a video?  |
|              | <p><b>Research Activity</b><br/>Design and iterative development of Videolyzer including heuristic evaluation and pilot user trials; Evaluation of Videolyzer with bloggers and journalists via questionnaires and interviews.</p> <p><b>Contributions</b><br/>In the process of using Videolyzer to do video analysis users' perceptions of comprehensiveness, various perspectives, context, and quality of the video were enhanced.</p>  |
| <b>RQ 3b</b> | Considering the novel annotation features implemented including transcript integration, granularity, and suggestive automation, what are the benefits and limitations of these features of the Videolyzer interface?  |
|              | <p><b>Research Activity</b><br/>Evaluation of Videolyzer via observation, questionnaires, interviews, log analysis, and screen recording.</p> <p><b>Contributions</b><br/>Transcript integration and coupling of video interactions to the transcript was both a usable and preferable method for navigating and annotating video. Granularity was perceived to help focus and organize the discussion around key pieces of the video or transcript. Automatically suggested annotations were of low to moderate benefit for aiding navigation and annotation.</p>  |
| <b>RQ 4</b>  | If the rich annotations collected using Videolyzer are presented as simplified visualizations to end consumers can the credibility of statements made in the video be modulated?  |
|              | <p><b>Research Activity</b><br/>Design of vastly simplified interactive visualization shown inline with a video to express cues such as activity and polarity of annotations as well as sources used. Evaluation of this visualization in a controlled online experiment administered on Amazon Mechanical Turk.</p> <p><b>Contributions</b><br/>We show that for some, but not all types of, content it is possible to module credibility evaluations using the cues we selected from Videolyzer annotations. We consider the limitations and implications for such visualizations in terms of user experience and credibility impact.</p>             |

## CHAPTER 2

### MULTIMEDIA ANNOTATION METHODS AND INTERFACES: RELATED RESEARCH

The value of annotated and metadata rich media, whether images, video, or text is well recognized in enhancing and easing information organization and retrieval. New forms of media applications can also be enabled when formal processing capabilities based on typed metadata are made available [9, 24, 26]. The primary problem is the cost and effort associated with acquiring additional descriptive or formal metadata. The purpose of this chapter is to review some of the existing work that has been done on reducing the costs associated with gathering annotations for multimedia data including images, video, text, and visualizations. There are at least three methods that have been explored in the literature for reducing the human cost associated with annotating media. They are (1) full automation, (2) reducing actual costs of manual annotation interfaces through usability, semi-automation, and collaboration, and (3) reducing the perceived costs of manual annotation.

#### **Fully Automated Annotation**

While the holy grail of media annotation is automatically generated, semantically rich metadata, the state of the art in computational perception is still somewhat limited to producing mostly low-, signal-level metadata and some higher level metadata in constrained contexts. Performance outside of constrained domains (and sometimes even within them) can be of limited accuracy and

usefulness. A recent review of content-based image annotation research [116] suggests that, while there have been recent gains in the percentage of words learned in automatic image annotation (up to about 50%), advances are slowing and may be asymptotic. Even for static objects there are fundamental issues associated with fully automatic annotation, notably the contextual and subjective interpretation of photos or images which may not be denoted by the pixels themselves. In other words, connoted meanings do not fully reside within the pixels present in the image and require further human-level interpretation within a larger symbolic or cultural system.

Automated annotation has been explored in the realm of video using visual, auditory, and textual streams of information. One of the canonical multimodal video analysis systems is the Infromedia project, which focused on the analysis of news video archives for the purpose of improving queries and information retrieval [124]. In that system, image processing was used to extract scene boundaries, faces, and text which could be run through Optical Character Recognition (OCR). Automatic Speech Recognition (ASR) was used on the auditory stream to produce a transcript which could be indexed, albeit the recognition accuracy was only 66%. Named entities were then recognized in the transcript and OCR extracted text and were correlated with faces. What makes the Infromedia system quite interesting is the degree to which the different annotation streams are fused, allowing metadata redundancy to mitigate some of the errors in any one type of metadata.

Researchers continue to work in this space [71], though they acknowledge the difficulty in overcoming the semantic gap [105], which is the difference between

the semantics of a medium as seen by a human and the low-level features that can be extracted by a computer. For instance, Davis [19] talks about the types of automated annotations of video that computers are getting good at: shot break detection, classification of pan and zoom, object tracking and recognition, and audio cue detection; most of these are fairly low-level features. Davis argues that video repurposing applications must be supported by human annotation, at least in the near term. In general, the success of these systems largely relies on the types and degree of semantic richness of the annotations needed to support the desired application. In the case of Informedia the names of people and places as correlated through the multimodal streams was quite helpful for their particular information retrieval application.

### **Manual and Semi-Automated Annotation Interfaces**

Considering the semantic gap and the nature of the annotations needed to support some applications it is natural to turn toward fully manual or sometimes semi-automated interfaces. One of the chief issues and a valid criticism of many manual annotation systems is that they are too costly in terms of user time and effort and do not engender usage [18, 31, 102]. The focus of much research in this area has been in reducing costs and time associated with adding tags or annotations by simplifying the interface or by including some automation to reduce the user burden, a flexibility principle in HCI known as task-migratability [30]. In the remainder of this section I'll look at related research in the video domain and to a lesser extent in the image and text domains, some of which has tried to reduce

annotation costs through (1) usable interface design, (2) semi-automation, and (3) distribution of work across collaborators.

## **Interface Design**

### Video Representational Confounds

Video is an unwieldy medium to annotate due to its time-varying nature across dual channels (visual and auditory). This leads to difficult representational and interactional issues in interface design relating to navigation, segmentation, and selection of the video in the process of manipulation [16, 25].

One of the representational issues involves scale and in providing the appropriate level of granularity for the task at hand; different granularities and scales are needed to support the range from shot or scene based annotation all the way down to frame accurate annotation [11]. The way in which selections and segmentations are made in the interface also hinges on the granularity chosen. Mills et al [83] addressed this issue with a hierarchical video magnifier, which allowed users to work with video at a high degree of detail, while maintaining an awareness of context. Using multiple hierarchical timelines, the user could successively drill down into a video, allowing for a detailed focus plus context visualization.

Ryall et al. developed the concept of the temporal magic lens [96] to address the issue of browsing, navigating, and querying video data. Temporal magic lenses define a spatial and temporal extent in the video which is extracted and composited in the interface to provide a summary. They also developed a timeline widget, aligned with the time of the video, which showed object frequencies as stacked bars.

Depicting additional metadata streams in the video as they become available can also facilitate analytic activity.

Yamamoto et al. [127] proposed a time-based visual presentation framework oriented towards facilitating active watching and analysis of video using a set of twenty seven interface primitives. Active watching is meant as an analogy to active reading, which involves combining reading with critical thinking and learning [15]. The different interface primitives address different transformations in the temporal or visual presentation of the video. Temporal transformations include slowing down, speeding up, reversing, or repeating. Visual transformations include things such as size, zoom, or rotation. The framework is meant as a generally applicable set of transformations that might be useful for different types of video analysis tasks.

### Video Annotation

One of the earliest interfaces in the literature to approach video annotation was Mackay's EVA [74]. EVA was a video analysis application which allowed for tagging of events in real-time and was developed to allow for behavioral video analysis. The user could add events of interest and link free text notes and tags in the audio track either in real-time or in the post-processing data analysis phase. The minimalist interface of EVA was meant to facilitate simple and low-cost annotation, although this came at the expense of providing for richer, more sophisticated analytic activity.

Marquee [125] was another early system that was initially developed for real-time video annotation and tagging using a pen-based computer. Design iterations showed that users also needed a mechanism for applying keywords

without the pressure of real-time however. Marquee used the notion of segmenting the video into timezones to which keyword “paint” could be applied by dragging the keyword vertically along the timezones to which it should be applied. The study of Marquee showed that overviews and representations are very important for making sense of annotated videos and that logging video is an iterative process, not something that can be completed in real-time in one pass.

The concept of video lenses was introduced by Costa in [16] as a mechanism for customizing the view or rendering of video metadata according to users’ tasks or interests. Different lenses (e.g. transcript, who is in the shot) could have different interaction metaphors to facilitate entering and rendering different types of metadata. In their VAnnotator system, a multi-layer timeline representation was used with each track or layer corresponding to a particular video lens.

The LEAN system presented by Ramos in [93] was another effort toward novel visualization and interaction techniques for video annotation specifically designed to be used with pressure sensitive digital tablets. The system strove to make video annotation more like annotation using freeform hand-writing on paper, which would presumably decrease annotation costs due to most peoples’ familiarity with pen and paper. However, because of the free form ink method, annotations are not machine readable without a handwriting recognition post-processing step and thus are not immediately useful for further indexing, aggregation, or processing using the computer.

Another recent attempt at a comprehensive video annotation interface is VCode along with its annotation data rectifier, VData [45]. The issues that Hagedorn

et al. address in VCode are: poor support for annotation workflow, representational issues of data on timelines, and inadequate interaction techniques with video, data, and annotations. The workflow engineering and inclusion of scheme design and reliability testing with VCode are notable integrations, which make the system smoother for the intended application of behavioral video annotation. At an interface level, they define both ranged and momentary timeline marks with the momentary marks compressed to a single sub-timeline and the ranged marks each having their own timeline. Interactions included a subset of those defined by Yamamoto [127], which they found to be relevant to the behavioral video researchers they had interviewed.

Another interesting aspect of VCode was the inclusion of contextual data streams which were visualized as bar, line, or scatter graphs aligned with the timeline of the video. In other video annotation and browsing systems such as VACA [10] and VAST MM [47] additional contextual information (via automatic content analysis) was used to add structure to the video thereby easing navigation for certain information finding tasks. A related strategy is implemented and evaluated in the Videolyzer system presented in Chapter 5.

### **Semi Automation**

One of the more promising avenues for annotation of media is to combine automated algorithms with the manual annotation efforts of people. Automated algorithms can work together with humans in at least three ways: (1) to highlight areas of media of potential interest for a human annotator thereby reducing human effort in navigation and browsing, (2) to suggest concepts or tags that might be

relevant annotations, and (3) to automatically annotate the media which the human annotator can then correct of any inaccuracies and false detections. In all these ways the automatic algorithm makes *suggestions*, which the human uses as a starting point for further annotation or correction.

In the text domain, the ClaimSpotter tool was designed for the sensemaking of scholarly papers and suggested (and highlighted) areas for annotation based on a document's "important" sentences and by detecting pertinent relations from an ontology [99]. Several semi-automatic strategies were developed for photo annotation in the EasyAlbum system [18]. Clustering and ranking algorithms were employed so as to reduce the amount of time users had to spend on photo navigation and selection before applying a tag. The Family Video Archive (FVA) explored the symbiosis between automated and manual techniques for tagging collections of videos [4]. FVA provided suggestions for structured tags based on free-form text that had been entered by the annotator.

Videotater also attempted to alleviate some of the tedium of video tagging on a pen computer by facilitating the rapid correction of automatically generated segments [25]. The segmentation task, a sub-task of tagging, was supported using a gesture that would snap to the nearest likely boundary between shots automatically using inter-frame color histogram comparisons. Refinement and correction of automatically produced segments was also enabled using a specially designed visualization which allowed the in and out points of an automatically detected segment to be refined quickly.

## Collaboration

Another method that has been used to reduce the individual effort involved with tagging photos or other online content is to distribute the work across many people. This is the concept behind social or collaborative tagging, which allows users to assign unrestricted keyword tags to content which is shared with others [72, 78, 98]. This model has been successfully employed by the community of users on Flickr [81] who are interested in the organizational and social aspects of tagging photos. In an early study of tagging behavior on Flickr Marlow et al [78] found that an “overwhelming majority” of tags on Flickr are added by individuals to their *own* photos and not to the photos of others. So, whilst social tagging incentivizes the annotation process with the hope of future organization, attention, or recognition, the promise of dramatically parallelizing the tagging process across large numbers of people is perhaps unmet.

Distributing the annotation of video across multiple people using collaborative software has also been examined by a few researchers. Cockburn and Dale developed a synchronous collaborative video annotation system, CEVA [14], which had the goal of enriching the video annotation process through either synergistic or parallel analysis. Synergistic analysis describes the additional insight gained from having multiple video analysts share their different views and perspectives. Parallel analysis is the idea of using multiple analysts to speed up annotation by distributing the workload.

Volkmer et al. [118] built the EVA system to address issues associated with the parallelization and distribution of a large scale video annotation task so as to

speed up the annotation process as much as possible. The annotation task used was somewhat restrictive allowing for only static concepts, pre-defined tags (i.e. there was a controlled vocabulary), and shot level granularity. Individual key frames were annotated and the annotation taken to hold for the entire shot. Customization of the interface in terms of annotation style, number of passes, mouse or keyboard, and layout was included so that individual annotators could setup their annotation environment in the way that was most efficient for them. The results of the user study suggested that tags would be generated more efficiently and with higher quality if annotators were only tasked with tagging one concept at a time.

Another approach toward collaborative annotation harnesses the wisdom of the crowds to collect and verify information about identification or categorization tasks where automatic algorithms typically perform poorly. Popularly referred to as crowdsourcing and academically referred to as human computation, this approach aggregates the repetitious annotations of many people in order to collect corroborated and validated annotations. Although the concept was first pioneered by Luis von Ahn, others have begun exploring this area as well. For instance, in the HumanSense project [128] a framework was developed to aggregate annotations from many independent contributors while filtering out spurious information from potential spam attacks or from human contributors trying to cheat the system or collude to taint the collected data. They also explain a multi-step image annotation application within their framework that takes initial annotations from contributors and subjects them to further refinement in subsequent steps of the human computation algorithm. The result is an increase in the information quality of the

collected annotations. The system is envisioned as analogous to the way advertisements are served on webpages; a webpage can embed a human computation task and then visitors to that webpage have to solve that task in order to gain access to the content on that site or to receive some credit or payment.

### **Blended Approaches: Semi-Automation and Collaboration**

Some of the most promising approaches toward rich annotation of media are hybrids, which employ both semi-automation and usable collaborative interfaces to enhance the information quality of the annotations collected. One instance of where this strategy has been successful is in opening up automatically produced transcriptions of video for collaborative editing using a Wiki format [86]. In this way the inaccuracies of an automatic transcription algorithm are corrected by many users contributing edits to the transcript. In Munteanu et al.'s work in this area they found that transcript Word Error Rates (WERs) were reduced by an average of 53% across a set of 21 videos used in their experiment. These reductions however come at the expense of other types of inconsistencies such as different wiki editors using abbreviations, formulas, names, or spelling errors. While these don't overly impact subjective readability, they do introduce other types of errors in the collected metadata. A further complication of the work concerns how to motivate people to contribute to the improvement of the transcript. The strategy used in Munteanu's study was to require students to edit the transcript in order to gain access to the video information, or to give students extra credit for the time they put in editing.

Another effort that combines automatic and manual methods for annotation is the reCAPTCHA system [123]. CAPTCHA, standing for "Completely Automated

Public Turing test to tell Computers and Humans Apart” is a web security method that presents a string of distorted characters that a human must type in correctly in order to pass. It is typically used to gain access to websites, new user accounts, or to post comments online. The assumption is that computers are unable to decipher the correct letters because of the distortion applied to the letters. reCAPTCHA is an extension of this idea, but instead of distorting known characters it asks users to type in words scanned in from old books. To pass the reCAPTCHA the person must type two words: one that the computer already knows the answer to (via OCR software) and another that the OCR software could not decipher. If the human answers the known puzzle correctly then the algorithm assumes the unknown word was also typed correctly. The person answering has no knowledge of which puzzle the computer knows the answer to. Using this mechanism, words that an OCR program could not recognize correctly are corrected one by one with human effort. Words are corroborated by multiple users independently to ensure valid and accurate transcriptions. The mechanism is tremendously successful, boasting transcription of over 440 million words with an accuracy exceeding 99% [123].

### **Perceived Costs of Annotation**

An alternative to engineering the user interface to objectively minimize actual annotation costs is to consider engineering costs of annotation as *perceived* by users. The greatest objection to manual tagging interfaces is that they are tedious and boring to use, but if they were engineered to be more enjoyable or fun [46, 76, 101], the perceived costs of usage would be reduced even if the actual time costs remained the same. On the other hand, users may also be persuaded to put more

effort and time into annotation if there was a greater perceived or actual usefulness for the annotations [63]. This is akin to the carrot and stick metaphor; by providing more interesting applications which rely on the metadata, the hope is that this would motivate more people to add the necessary metadata.

There are many potential strategies for reducing perceived costs by providing additional incentives to use annotation or tagging systems, including both intrinsic and extrinsic motivators. Extrinsic motivators include things such as money, prizes, social adulation, or somatic rewards (e.g. food, drugs, sex, or humor [84]) and are external rewards for behavior. On the other hand, intrinsic motivators include such things as providing for a sense of autonomy, competence, or social connectedness [20]. Many forms of activity can become intrinsically motivating if they induce what psychologist Mihaly Csikszentmihalyi terms flow [17], a mental state in which a person is fully immersed in what they are doing. Flow involves intense concentration, clear goals and feedback, and a balance between the ability level of the person and the challenge of the activity. Ideally, we would want to use any and all intrinsic motivating strategies because they are self-sufficing and don't require a constant infusion of external resources to keep the behavior happening.

Several recent studies have looked at the incentives and motives that people report for participating in tagging activity online. Research looking at photo tagging behavior on Flickr [78, 89] has suggested that incentives for tagging there are both in organizing one's own photos and in the opportunities to share with one's contacts. An interview study of people participating in tagging on Flickr [5] was used to induce a taxonomy of motivations including dimensions of functionality and

sociality. Functionally, they found people tagged either for their own organization or for the purpose of communicating with themselves (in the future) or with others in their friends and family network. In a study by Sen et al. [98] they showed empirically that more users were converted to taggers on a movie website when their tags were shared with others versus when they were unshared. In social tagging applications, awareness and visibility of tagging seems to provide some impetus for tagging activity. These studies suggest that tagging activity in online communities is largely related to intrinsic motivators such as social connectedness.

### **Annotation Via Lusory Interfaces**

On one hand you have human computation approaches like reCAPTCHA and HumanSense which are based on external motivations such as money or access to resources. But another way to achieve human computation is to motivate it intrinsically. A reframing of the task of annotation from a work task to a play task, such as a game, fundamentally alters the perceived usage cost by users. This has been explored by von Ahn in a series of games that were designed to make use of human computation to annotate images or collect other useful data for machine learning applications [119-121]. In the original work on the ESP Game [119], pairs of unacquainted players were shown the same image and asked to guess what their partner was thinking about the image. Since the players could not collude or otherwise communicate, the guessed words ended up being good tags for the images. When multiple (in this case two) people independently assign the same word to the same image, this validates the word as a good description of the image.

The ESP Game leverages a game mechanic similar to Taboo, a word game in which players have to describe concepts without using a list of taboo words.

Follow on work by von Ahn extended the use of games and human computation to collect captions for images in Phetch [120] and to segment objects in images in Peekaboom [122]. Von Ahn's work stresses the value of validating the correctness of annotations in his games but does little to critically evaluate the enjoyability of the games aside from citing usage statistics. More rigorous attention to the playability and enjoyability of games would lend more credibility to claims about annotation using lusive interfaces [21, 113].

Another photo tagging game, PhotoPlay [22], was designed with the intent of combining incentive structures to provide both a game environment and a social milieu with opportunities for social engagement and sharing of one's photos and stories with collocated friends. PhotoPlay differs from the ESP Game in several key respects; most importantly it is a collocated game for up to four players, which instead of stifling communication between players, is actively designed to promote it. The system of tag validation is also consensus based, but takes into account several human perspectives in negotiating consensus. Also, PhotoPlay has a provision for playing with personal photos in addition to online photos, which facilitates a degree of social engagement not available in the ESP Game.

Recently, the notion of using games to motivate annotation of media has also been extended to the medium of audio and music by Turnbull et al. in the Listen Game [115]. The game is very much inspired by the ESP Game but makes a foray into the realm of collecting tags for musical snippets that are more subjective and

opinion oriented. Instead of collecting binary information about a label, the game collects information about the strength of the association between a word and a song as aggregated across a number of subjective opinions. The game works by presenting players with a 15 second long snippet of music and a set of 6 words from which the player must choose the best and worst words to describe the music. The player earns points based on the amount of agreement between their choices and the choices of all other players. It remains unclear that this is a good way to do scoring since it still incentivizes consensus rather than divergence between players.

### **Summary**

There is a wealth of research that has been done in both image and video annotation interfaces to reduce the costs of annotation through automation, usability, and collaboration. Video in particular is a difficult medium because the navigation and selection needed for annotation are more cumbersome in a time-based medium. An ideal strategy for reducing the costs associated with collecting rich semantic metadata should take advantage of as much automation as is feasible, should provide for a usable interface, should incorporate collaboration to distribute the task and improve information quality, and should motivate people to contribute and tag content. There has also been recent success in manipulating the perceived costs of annotation by reframing the task as a lusory one, which can provide enjoyment and fun while also collecting useful metadata.

In the following chapter we detail the work undertaken to design and evaluate a game, Audio Puzzler, for the annotation of time-stamped speech transcripts of video. Chapter 4 and 5 describe the design, implementation, and

evaluation of a more general video analytic annotation system, Videolyzer, which incorporates aspects of semi-automation and collaboration in its design strategy.

## CHAPTER 3

### ANNOTATION OF AUDIO-VISUAL MEDIA WITH TIME-STAMPED TRANSCRIPTS USING A GAME

Audio Puzzler is a casual puzzle game which uses real spoken audio clips taken from short videos as puzzle pieces [28]. Players must first transcribe and then piece together snippets of these audio clips in order to complete the puzzle. Audio Puzzler is designed to be fun and enjoyable to play while also producing meaningful output in the form of time-stamped transcriptions of the audio used for the puzzle. As suggested by the related work on tagging of images and music using human computation, we designed Audio Puzzler to manipulate the perceived costs of annotation, thus lowering the barrier to user participation in the activity.

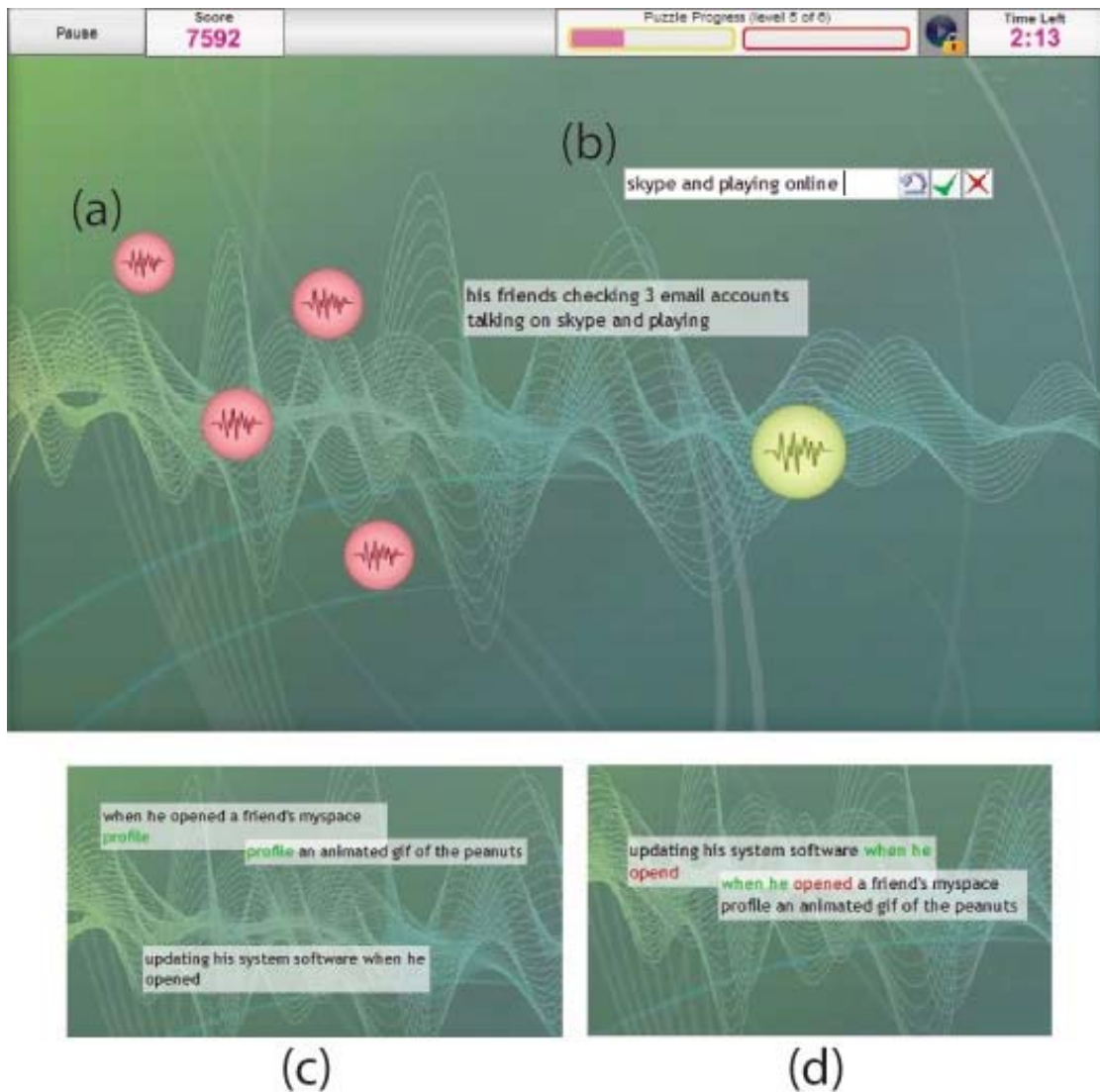
Accurate transcriptions are valuable for improving information access, searchability, and browsability of audio or video clips [87, 109], but current Automatic Speech Recognition (ASR) systems typically produce Word Error Rates (WER) of 20% to 45% in real world conditions [87]. Only in the best of acoustic conditions and with training for a particular speaker can the WER get much lower. Furthermore, computing accurate time stamps for individual words from imperfect ASR transcripts is an area of current research in multimedia [48]. Having precisely aligned transcripts enables new types of multimedia interfaces which tightly couple the transcript to the timeline of a video.

Our system approaches the time-stamped speech transcription problem using a luserory (game-based) human computation method. We evaluated our system in the laboratory (1) by conducting a user study to assess the enjoyability of the

game aspects and (2) by comparing the transcriptions and time-stamps generated using the game to manually labeled ground truth transcripts. In a follow-up study, we made some design changes to the game and put it online with a wider range of content for people to play with. Using the data collected from that study we developed an aggregation algorithm which improved the accuracy of the transcripts produced even further. In the following sections we present the design, algorithms, and evaluation of the system both in the laboratory and in the field.

### **Audio Puzzler Game Design**

Audio Puzzler is a Flash-based online puzzle game that uses speech audio as the basis for the puzzle pieces. The goal of the game is to assemble the puzzle pieces as quickly and accurately as possible. At the outset, each puzzle starts as a set of audio bubbles (Figure 3.1a), which are grouped by color to simplify the puzzle. Double-clicking an audio bubble pops the bubble and plays the audio in that bubble. The interface then allows the player to type the words that were spoken in the audio (Figure 3.1b). Audio can be repeated by clicking an icon or pressing a key combination. Once typed, the textual puzzle pieces can be dragged on top of one another to connect them. If the text matches and they belong next to each other, then the matching words are highlighted in green (Figure 3.1c). Dropping the piece merges them. If, for some reason, the words should match but do not (e.g., because of a typographical error), then they are highlighted in red (Figure 3.1d). As pieces are assembled, progress is shown at the top by filling in the colored meters. Time and score as well as a pause button are also shown in the top area. Upon completing



**Figure 3.1. Overview of the Audio Puzzler Interface.**

the puzzle, the player can optionally listen to and watch the video from which the audio was taken.

Each audio puzzle consists of 3 levels of increasing difficulty, each with more puzzle pieces to assemble. The amount of audio in each level roughly corresponds to 15s, 20s, and 25s. For each level, the audio is broken into a number of overlapping audio chunks represented as bubbles. The overlap is essential so that the player can see where the puzzle pieces match up once they are typed. Bubbles containing only silence (according to a simple short-term energy feature) are filtered out of a level.

The scoring system in Audio Puzzler is designed to promote quick completion of the puzzles and to make the game more challenging by providing a sense of time pressure. The game clock starts off with 3 minutes and counts down. Each time the player successfully merges two pieces they get more time on the clock. The amount of time added is about 17 seconds and was determined through early play testing so as not to make the game too easy or difficult. Also, when the player merges two pieces they get a number of points proportional to how early in the level the merge took place. Merges earlier in the level receive more points than merges later on. This encourages connecting the pieces and completing the puzzle as quickly as possible. Additionally, bonus points are awarded at the end of the level based on how quickly the level was completed. If the time clock falls below zero, the game continues (as completion produces more transcription), but the player is penalized 5 points every 5 seconds until the clock rises above zero again. This, too, lends a sense of pressure and urgency to the game. At the end of the game, the player receives another bonus (or penalty) based on the time left on the clock.

Care was taken in the design of Audio Puzzler to adhere to good heuristics of game design such as providing adequate feedback, control, and cognitive engagement [21, 59, 90]. Animation and color are used throughout the game to make it more “juicy” to play [101]. Bubbles bob and pulse, hearts shoot from merged pieces, and large text animations mark the completion of color groups and levels. We also tried to balance the fun of the game against the underlying goal of generating good transcripts by setting parameters such as chunk length, overlap, and game length.

## Transcription Algorithm

Before text from audio chunks is merged to form the transcript, it is normalized using a set of filters so that comparisons to the ground truth are more fair [51]. Contracted words such as “*we’re*,” “*They’re*” etc. are expanded to “*we are*,” and “*they are*”, respectively. Also, numbers written with numerals are expanded so that, for example, “*2008*” becomes “*two thousand and eight*.”

At the completion of a game, we have a set of  $n$  audio chunks  $C_1 \dots C_n$  with their transcriptions ordered by start time. Chunk duration is chosen randomly at the start of the game to be 2.25s – 2.50s and each chunk has a .75s overlap with the chunk preceding it. These parameters were chosen to facilitate the user only having to listen to each chunk once during game play as well as to ensure sufficient word overlap between pieces. By looking at the average spoken word rate for English (about 160 words per minute or about 2.67 words per second [131] in conjunction with the short term memory limits of humans (about 7 chunks [82]), we determined that an audio chunk length of about 2.5 seconds would lead to about 7 words per chunk. Thus, the player should be able to keep these 7 words in short term memory during transcription and avoid having to replay the audio.

To merge the audio chunks we take each pair of adjacent chunks  $C_i$  and  $C_{i+1}$ , for  $i = 1$  to  $n-1$ , and convolve the words of  $C_i$ , with  $C_{i+1}$ . If two words match then the score at that point in the convolution is incremented. The highest score is taken as the optimal overlap point. The words before this point are taken from  $C_i$  and the words after this point are taken from  $C_{i+1}$  as shown in Figure 3.2. The best overlap point is computed with,

$$\text{Max index} \left[ \sum_{j=0}^{|C_i|} \sum_{k=0}^{|C_{i+1}|} C_i^{j+k} \cdot C_{i+1}^k \right]$$

To compute time stamps for each word, we make the assumption that over the course of a chunk, words and pauses occur at a steady rate. Better approximations to the duration of each word and to inter-word pauses could be made using TTS (Text to Speech) techniques. Since we know the beginning and ending time of each audio chunk, each word is assigned a beginning and ending time by equally dividing the chunk duration by the number of words in the chunk taking into account inter-word pause duration, which we set at 20ms. For words that occur in multiple chunks (e.g., the last two words in  $C_i$  shown in Figure 3.1) the time-stamps are estimated and averaged from both chunks. If the end time of any word follows the start time of any word, we assign the average of the two as the end of one and the start of the other. This ensures that word time stamps do not overlap.

### Laboratory Evaluation

Our high-level measure of success was a game that is fun to play and also produces valuable transcripts. To evaluate Audio Puzzler, we undertook both (1) a



Figure 3.2. Merging two audio chunks. Optimal overlap point shown in solid red.

user study to assess aspects of the enjoyment and usability of the game and (2) an analysis of the accuracy of transcripts and time-stamps output from the game.

### **Audio Puzzler Interface**

Based on game design heuristics, the user study was developed to assess the game's enjoyability and fun, challenge, replayability, pacing, and usability in terms of feedback, game status, and controls [21, 76, 90]. We first ran a pilot study with 5 participants to understand most of the usability issues and iterate on the interface until it was stable and usable. We then ran the main study.

Playtesting of Audio Puzzler involved 3 methods: observation, questionnaires, and logging. Observation of players was used to identify reactions such as smiles and laughter, cognitive difficulties or confusions with the controls, and content effects such as indications of engagement with or comprehension of the audio content. A questionnaire was used to collect self-reported information about participant background, fun, replayability, usability, pacing, and challenge. User scores, elapsed time, and the transcriptions were also logged for analysis.

Ten people participated in the main user study (9 male, 1 female). Eight participants were native speakers of English and 2 were not. We first explained the instructions for the game and briefly demonstrated the controls to participants. Next, we asked each participant to play the game twice, once each with a different audio puzzle. In total there were 3 puzzles used in the playtesting; each participant played with 2 out of the 3. The order of the presentation of the puzzles was counterbalanced across participants. After the second game, participants filled out the questionnaire and were informally interviewed.

The content of the three audio puzzles was selected for understandability from a set of five video clips used in the pilot. We found in the pilot that if the speaker in the video clip used awkward grammar or spoke too slowly this affected the playability of the game. The final content consisted of a news clip about medical error, a snippet of a speech made by Al Gore taken from *An Inconvenient Truth*, and a portion of a fake comedic newscast from *The Onion*. Each clip was 60s to 70s long.

Summary statistics for fun, challenge, and pacing from the questionnaire and game time from the log analysis are shown in Table 3.2. The numbers show that participants found the game to be fun (4.65), but too challenging (3.3) and too long (5.75). Some participants laughed and smiled while playing the game. Others were vocal, making comments like, “yay,” “*that was fun*,” or “*yes, ohh boy*” after completing a level or game, or just finding matching pieces. Several participants said that the idea of a puzzle involving audio was interesting and that trying to figure out how to put the puzzle pieces together was engaging and fun. One wrote, “*I liked the challenge of anticipating where different clips were in the story line.*” A few participants mentioned competition and getting the high score as another fun aspect.

Content was a significant component of what was fun or not about the game, with participants indicating that the two informational clips, while informative, were less fun to play with than the *Onion* clip. Although this is supported by quantitative responses on the questionnaire, the difference was small and statistically insignificant; the average fun for the informative clips was 4.62 and the average fun for the *onion* clips was 4.71. Others liked the challenge of multitasking

(i.e., listening, typing, and scanning for matches) or the challenge of the time pressure. Several participants mentioned that they liked the content they played with whether that be informative or entertaining; one participant appreciated being able to learn from the content of the game. Among the things that participants disliked about the game were that a social element (either competitive or cooperative) was missing and that there was too much typing involved in the game.

**Table 3.2. Summary Statistics for Game Play Ratings (7 points scale).**

|  | <b>Average</b> | <b>Std Dev</b> |
|--|----------------|----------------|
| Fun (1 not fun, 7 fun)                   | 4.65           | 1.18           |
| Challenge (1 too hard, 7 too easy)       | 3.3            | 1.06           |
| Level Duration (1 too short, 7 too long) | 5              | .48            |
| Game Duration (1 too short, 7 too long)  | 5.43           | 1.25           |
| Game time (seconds)                      | 875 s          | 23.90 s        |

Some players commented that spelling was an issue. It was sometimes difficult to hear words because they were truncated at the beginning or ending of a chunk, and the game was more difficult if you were not a good typist. One participant thought it was too cognitively demanding to play, writing, *“It took a lot to listen to the audio and then type it and also think in the back of my mind about what I’ve heard so that I can make connections as soon as I was done typing the audio.”*

During our observations of game play, we took special note of the effect of time pressure. We found that some participants appeared unaffected by the time pressure and enjoyed the game for the sake of solving the puzzle. Others were fixated on time and with being the fastest and best. This mentality led to play strategies which involved the use of abbreviations, numerical representations of numbers, and the elision of some words for each piece to reduce typing. These

participants saw that there were usually multiple words overlapping between adjacent audio pieces and learned that they only needed one overlapping word to make a match. One participant went so far as to try transcribing only the first and last words of each piece, but soon found that this was ineffective for rapidly finding matches and reverted to typing out most words.

We observed how important context is for understanding individual words and for transcription in general. In several cases, participants had trouble understanding particular words or phrasing of isolated pieces. For example, a participant first transcribed a word as “gift” but upon hearing the context of the word when another piece was played corrected the transcription to “animated .gif.” Participants would often go back and correct transcriptions of pieces as they heard more audio pieces and better understood the context of previous pieces.

### **Transcription Accuracy**

The second goal of our evaluation was to analyze the accuracy of the transcripts and time stamps produced. We created a ground truth transcript with time stamps of each word for the 3 audio puzzles. We manually marked the beginning and ending time of each word in the clip, a tedious process requiring about 3 hours to transcribe 1 minute of audio.

We compared each transcript produced by a participant using the game to (1) the ground truth and (2) automatically generated transcripts produced using the Sphinx 3 speech recognizer. We computed the Word Error Rate (WER) in comparison to both the ground truth and ASR output [51]. We also computed the

mean error for the time stamps of words that were aligned between the hypothesis transcript and the reference transcript using the following equation,

$$\sum_{T_i=\hat{T}_j} .5 \cdot \left( \left| S_i - \hat{S}_j \right| + \left| E_i - \hat{E}_j \right| \right)$$

Where  $S_i$  is the start time of the  $i^{\text{th}}$  word in the hypothesis transcript and  $\hat{S}_j$  is the start time of the  $j^{\text{th}}$  word in the reference transcript;  $E_i$  and  $\hat{E}_j$  refer to the time stamps for the end of the word.  $T_i$  refers to the  $i^{\text{th}}$  word in the hypothesis transcript and  $\hat{T}_j$  refers to the  $j^{\text{th}}$  word in the reference transcript.

Summary statistics for WER and time stamp accuracies are shown in Table 3.3. The average WER rate for the three audio clips used was 10.08%, much lower than the 53.76% WER resulting from the ASR transcripts. The word time stamps generated using the game output were on average within .373s of the actual time-stamp from the ground truth. Based on the amount of time participants took to produce a transcript, we also computed the average play time need to produce each word as 4.95s.

**Table 3.3. Summary Statistics for Transcription Accuracy.**

|   | <b>Average</b> | <b>Std Dev</b> |
|---|----------------|----------------|
| WER (game output vs. ground truth)        | 10.08%         | 6.60           |
| WER (ASR vs. ground truth)                | 53.76%         | 10.85          |
| Time stamp (game output vs. ground truth) | .373 s         | .108 s         |

## Discussion

Literature suggests that even a 25% WER for a transcript is acceptable to start enhancing information searching and seeking behavior in multimedia systems

[87, 109]. Thus, Audio Puzzler's WER of 10.08% can be viewed as a success.

Informal analysis of the transcripts revealed that many of the errors produced using Audio Puzzler are the result of misspellings or typographical errors. This suggests that WER results can be further improved by running transcripts through a spell checker. In the next section we will detail the results of an experiment where we fuse the data from multiple players of the game to aggregate even better transcripts. When a group of people play the game with the same puzzle, correlations lead to even better WER results.

Another factor that affects transcription accuracy is the background, language ability, and knowledge of the player. There was a negative correlation,  $r(8) = -.55, p = .10$ , between the number of years our participants had lived in an English speaking country and the WER of their transcriptions. While some of this effect is clearly due to language ability, another part of it may be cultural knowledge. In the Onion clip there were many references to popular culture (e.g., bands, singers, products, etc.) that may have been more difficult for players to recognize and transcribe if they did not have cultural familiarity with those named entities.

We found Audio Puzzler to be moderately fun for participants based on the puzzle aspect of the game, competition, and scoring. There was a strong self-reported content effect and when asked if they would play the game again, several participants mentioned that content (and choice of content) would be a key determinant. While comedic content appealed more than informative content in general, participants suggested all sorts of different content that would appeal to them individually, including song lyrics, standup comedy, or soft news. A couple of

participants mentioned that they would be more motivated to play the game if they played with personally relevant content, such as political speeches, knowing that the output of the game would be a transcription of the speech from which others might benefit.

### **Online Deployment and Evaluation**

In order to collect more data on the accuracy of the transcripts that are produced with Audio Puzzler we made it available by putting it online at <http://www.audiopuzzler.com>. We included a wider array of content than in the initial study including news clips, sitcom snippets, and advertisements. We were interested in what types of people would come to play the game as well as how often they would complete puzzles and what the accuracy of the transcripts was outside of a laboratory setting. Between August, 2008 and January, 2009 over 2000 people played Audio Puzzler online. Many of the puzzles on the site have been completed by dozens of people, including one that was completed over 50 times. Because we were able to collect much more data, we were also able to begin experimenting with the accuracy improvements of transcript aggregation in order to correlate out errors.

### **Game Modifications**

Before we put Audio Puzzler online we took some of the feedback from the laboratory evaluation into account and tweaked the game design. Based on the user ratings, the game felt too long and was perhaps too difficult. To address this we modified each puzzle to consist of a greater number of shorter levels. The thought

was that this would lead to a quicker feeling of advancement and perhaps make the game not feel as long. We also added in bonus game pieces which did not contain audio but which when popped would automatically merge two other pieces that had already been transcribed. This was meant as a mini-reward that would save a little bit of time and sometimes make it easier to find matching pieces.

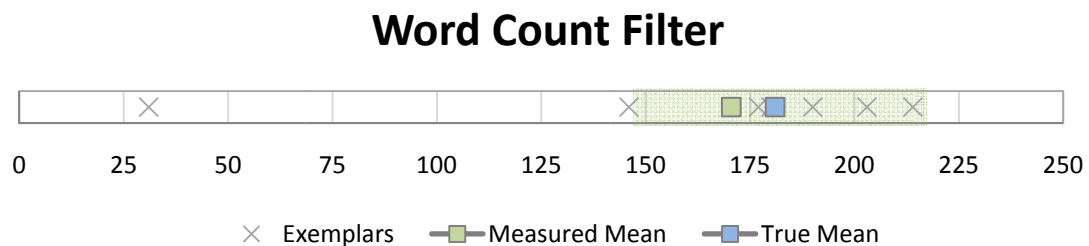
One comment made by several of the participants of the laboratory evaluation was that hearing partial words at the beginning or ending of audio pieces was confusing and added to the difficulty of the game. In order to make the listening component of the game easier for players and so that truncated words could be avoided in audio pieces we made an attempt at detecting word boundaries. This was done using a simple threshold on the Mean Square energy for each audio sample in a piece. A sample below the threshold was considered “quiet” and therefore could be classified as between spoken words. As each chunk of audio was turned into a piece for the game the initial endpoints were adjusted (within an acceptable range) so that they fell into a quiet area of audio.

### **Aggregation Algorithm**

We developed an aggregation technique in order to compile the best transcript from the many completions of a single puzzle (exemplars). Since it is impossible to compute an “average” of all exemplars collected (this would be meaningless), aggregation must produce a single transcript that takes into account the information from all exemplars. Aggregation can correlate out errors such as typographical mistakes and misspellings when enough exemplars are available.

The first step in aggregation is to filter out any deviant exemplars of obviously poor quality. This is accomplished by computing the mean and standard deviation of the number of words in each exemplar. Any exemplars with a number of words greater than one standard deviation away from the mean are filtered out. Also, any exemplars with a number of words less than one half a standard deviation away from the mean are filtered out. The reason for this asymmetry is that we assume that the true mean for the correct number of words in the transcript lies above our estimate. This is in turn based on the assumption that people are more likely to cheat by leaving words out (less typing and less effort) than by typing more words than necessary. In practice 28% of exemplars were removed from our corpora using this method. The concept is illustrated with real data in Figure 3.3.

Once exemplars of obviously poor quality are filtered out, we choose what we call the *buoy* transcript. Since we are unable to compute the mean of a set of transcripts, we select the buoy as the best median exemplar of the set. This is done by identifying the transcript that minimizes the WER metric to all other exemplars in the corpus. This is conceptually illustrated in Figure 3.4.

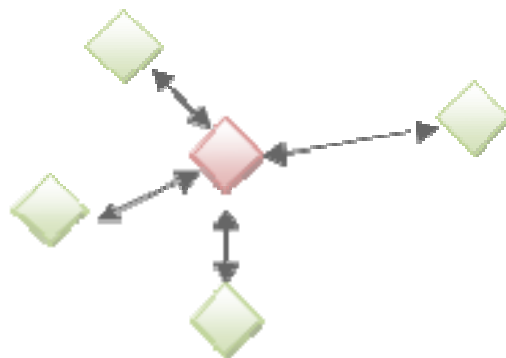


**Figure 3.3. Shows the true and measured means for the number of words recorded in a real set of exemplars. The light shaded range shows the area where the exemplars were retained for further processing.**

The buoy conceptually represents the most accurate transcript identified in the corpus, but it still may not be perfect. It may contain typographical errors, misspellings, or incorrectly recognized words. These errors can however be corrected using the information from the other exemplars in the set. This is done by a process of selective modification of the buoy including deletion, insertion, and substitution of words. We call this process *buoy boosting*.

In the WER algorithm used to compare two transcripts each word in the reference and hypothesis transcripts is marked in a different way: matched, deleted, inserted, or substituted. If two words in the reference and hypothesis transcripts can be aligned using the dynamic programming algorithm then they are marked as matching. Otherwise, words in the reference transcript can be marked for deletion or substitution. Words in the hypothesis transcript can also be marked for insertion into the reference transcript. We use these markings in the buoy boosting algorithm to iteratively suggest modifications to the buoy transcript.

In the deletion stage of buoy boosting words in the buoy with little corroboration in the set are removed. Each word in the buoy is assigned a



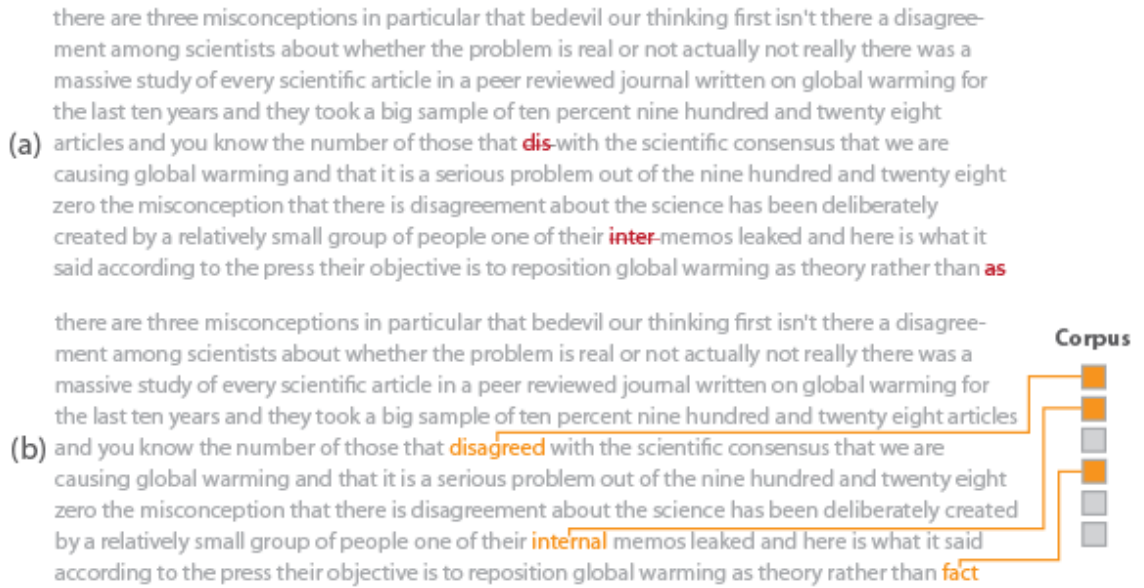
**Figure 3.4.** A buoy (red) is selected from a set of five transcripts by minimizing the distance between the buoy and the other transcripts based on the WER metric.

corroboration score based on how many times it is marked as matched when compared to all other transcripts in the corpus. If there are only three transcripts in the corpus (the minimum for running the aggregation algorithm), a word must have at least one corroboration in the set or it will be removed. If there are more than three transcripts, we require that a word have two corroborations or more or it will be removed. In other words, any words in the buoy that are not corroborated by at least one or two other transcripts in the set are deleted as anomalies.

Now we give each transcript in the set an opportunity to influence the buoy in a positive way by making word insertions and substitutions into the buoy. For each transcript in the set, the WER algorithm is used to identify words marked for insertion into the buoy. If a marked word has a corroboration score in the corpus less than half of the number of transcripts in the corpus then it is discarded, since inserting it into the buoy would then cause the buoy to disagree with more than half of the corpus. Otherwise the word is inserted into the buoy transcript and if this improves the average WER score of the buoy compared to the rest of the corpus then the insertion is accepted into the buoy. The same process is repeated for words marked as substitutions. The deletion and insertion process is shown with real data in Figure 3.5.

### **Transcription Accuracy**

Using the online version of the game we collected enough data to run the aggregation algorithm on six different puzzles including: an ABC News podcast about medical error, a snippet from the Seinfeld episode about Jerry being an anti-dentite, a political advertisement called Pump, an advertisement from the



**Figure 3.5. From a corpus of seven transcripts, (a) the buoy is chosen and words deleted which don't have corroboration from the other transcripts. In (b) words from several other transcripts in the corpus are inserted in the buoy to improve its accuracy.**

American Petroleum Industry (API), a snippet from an Onion newscast, and a snippet from the The Inconvenient Truth. The audio files, their duration, and the number of exemplars collected and remaining after the word count filter was applied are shown in Table 3.4.

For each of these cases we developed ground truth transcriptions for all further comparisons. We evaluated the performance of our aggregation algorithm by (1) computing the mean WER for the corpus against the ground truth, (2) computing the WER for the buoy versus the ground truth, (3) computing the WER for the boosted buoy versus the ground truth, and (4) comparing transcriptions

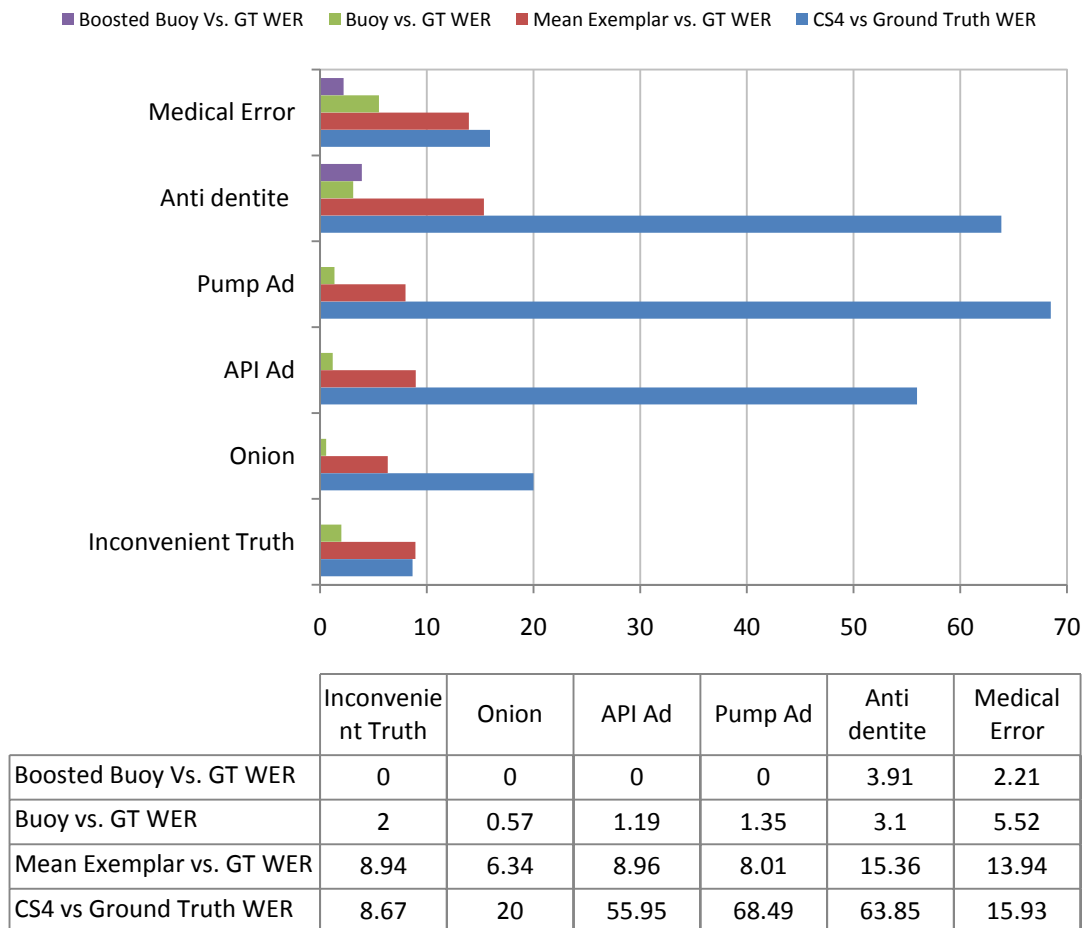
**Table 3.4. Audio Puzzles Used for Online Experiment**

|                            | Inconvenient Truth | Onion | API Ad | Pump Ad | Anti dentite | Medical Error |
|----------------------------|--------------------|-------|--------|---------|--------------|---------------|
| <b>Length (seconds)</b>    | 56                 | 61    | 30     | 30      | 51           | 71            |
| <b>Exemplars Collected</b> | 9                  | 13    | 25     | 49      | 10           | 12            |
| <b>Exemplars Remaining</b> | 7                  | 9     | 22     | 42      | 3            | 10            |

producing using Adobe Premiere CS4 to the ground truth. The results of these computations are shown in Figure 3.6.

Very similarly to the laboratory experiment, the mean exemplar versus GT WER was 10.26% (as compared to 10.08% for the lab data). Thus we can say that the quality of the transcripts produced in a real, ecologically valid environment (online) are of comparable quality to those produced in the laboratory.

Across all six cases the aggregation results show that the average buoy versus ground truth (GT) WER was 2.29% and the average boosted buoy versus GT WER was 1.02%. The aggregation algorithm boosted the overall accuracy by a



**Figure 3.6. Aggregation algorithm results showing WER comparisons to ground truth for each case.**

modest 1.27% on average. For four of the cases the boosted buoy versus ground truth WER was 0.0%, which means those transcripts were perfectly in concordance with the ground truth. These results indicate that extremely high accuracy (98%-99%) can be achieved with this method of buoy identification and aggregation.

Three of the audio clips used had minimal to no background noise and three had variable amounts of music, laughter, and noise in the background. The average CS4 versus GT WER for all six clips was an unimpressive 38.82%. But among the videos that were “clean,” that is the ones that had little to no background noise or music the average CS4 versus GT WER was a much improved 14.87%. Variable amounts of noise and the difficulty of neologisms wreak havoc on automatic transcription systems. But Audio Puzzler excels with these types of errors since real people are doing the transcribing. In either case, the method employed by Audio Puzzler is vastly superior in terms of the accuracy achieved.

### Sources of Error

In one case, Antidentite, the boosting algorithm hurt the final score since it deleted a word from the buoy that was not corroborated by the other two exemplars, but which was nonetheless correct. Two transcripts of lower quality can be detrimental to the initially superior buoy. Rerunning the aggregation algorithm with all of the exemplars (including those that were initially filtered out in the word count filter) resulted in a final buoy versus GT WER of 0.0%. This serves as a warning that with only three exemplars, boosting is more sensitive.

The remaining errors in the Medical Error clip were of three types: hearing, laziness, and hyphenation of compound words. In one error, many people

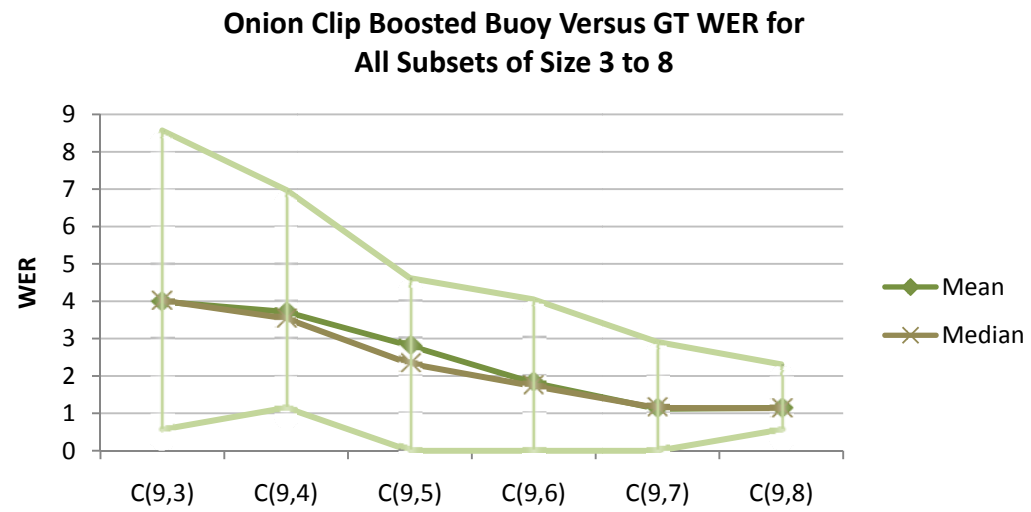
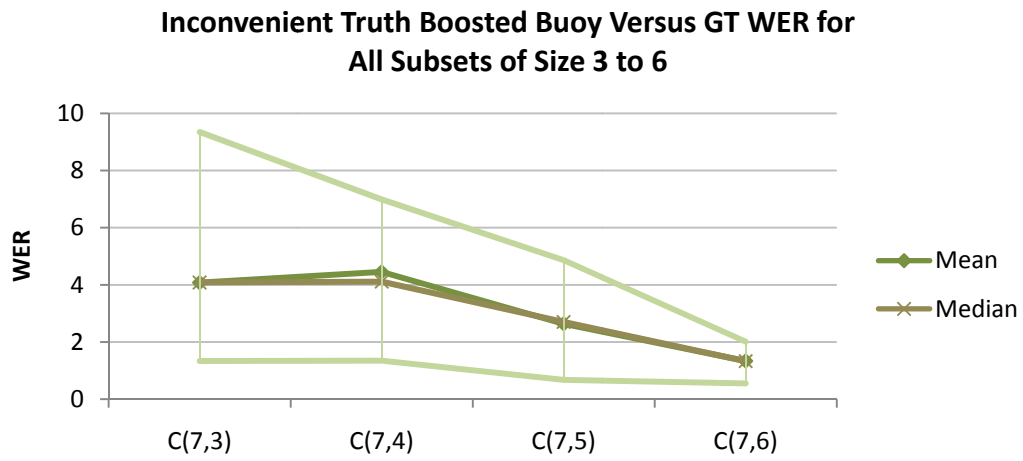
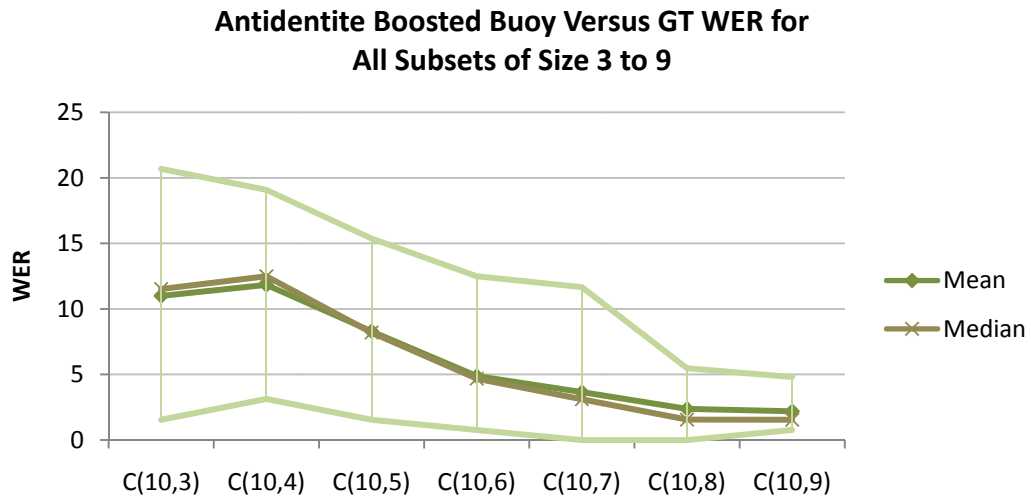
transcribed a “the” as an “a,” which was arguably hard to hear. An example of a laziness error is the elision of characters such as an apostrophe: “patients” instead of “patient’s.” Finally, for some compound words it is unclear whether they should be written as two words, one word, or two words with a hyphen. For instance, in Medical Error all of these variants were observed in the stenography of “mixups.” This is reasonable since words are learned through their use and the normative spelling may not be widely known by the population.

### Corpus Subset Aggregation Accuracy

To help characterize the effect of corpus size on the boosted buoy versus GT WER we ran the aggregation algorithm on every subset of transcripts from the corpora collected for three clips: Antidentite, Inconvenient Truth, and Onion. We only used these three clips because it would have been computationally intractable to compute the aggregated WER for all subsets of transcripts for the others. Results are shown in Figure 3.7.

As can be seen in the graphs, WER decreases as more transcripts are used in the aggregation process. As more information is fused from more transcripts the accuracy of the transcripts increases. In both Inconvenient Truth and Onion clips 5 transcripts was sufficient to achieve a worst case of no more than 5% WER, with medians in the 2-4% range. In the Antidentite case 8 exemplars were needed to achieve the same accuracy bounds. In Antidentite, Inconvenient Truth, and Onion clips, 10, 7, and 9 exemplars were respectively needed to achieve perfect 0.0% WER transcriptions. Some of the additional error in Antidentite can be attributed to

variations in spelling of names and stenography of “antidentite,” which needed more exemplars aggregated to arrive at a good consensus. The reason Antidentite has



**Figure 3.7. Aggregation Subset Results**

higher WER on the smaller subsets is that actually 7 of the transcripts from the corpus were filtered out by the word count filter. This test was run also to see how the aggregation algorithm would fare without the word count filter; with all 10 of the exemplars collected a perfect transcript was achieved.

## **Discussion**

Audio Puzzler is an effective method for gathering highly accurate time-stamped transcriptions of spoken audio or video media. With enough independent transcriptions, usually in the range of 5 to 8, our aggregation algorithm achieves a 2 to 4% WER. In this way we leverage many people indirectly collaborating with each other to enhance the overall data quality of the transcript metadata.

Our transcription accuracy is also high enough to compete with industry standard accuracy levels, which are about 98%. In the video transcription industry cheap labor is typically employed to produce time-stamped transcripts of video in three passes. The first pass is the initial transcription, the second pass aligns text chunks to the time of the video, and a final pass ensures quality control and pushes the accuracy up to about 98%. Automatic methods can only be employed in applications where lower accuracy levels are acceptable or the speaker and environment can be tightly controlled.

One drawback of both the automatic methods and Audio Puzzler as it is currently implemented is a lack of punctuation in the final transcript. In principle this could be added to Audio Puzzler, but in order to simplify the game punctuation was stripped. Punctuation could be added back in as a post-processing step if needed or desired.

## **Content Limitations**

What makes it difficult for Audio Puzzler, or any crowd-powered service for that matter, to be applicable for a wide array of content is that there is a need for privacy or confidentiality of information in many cases. For instance, a company may need video material from a customer interview or focus group transcribed. This video however contains confidential and proprietary information that makes it difficult if not impossible to make it available in a game that is free for anyone online to play. This limits the content that is feasible to transcribe with Audio Puzzler to audio and video that is already published online. This is no small amount of video, and of course includes news and information videos, but it does limit the overall applicability of the technology.

The content effect observed in the lab study also puts limitations on the type of content that can be effectively transcribed with the system. Engaging content is more fun to transcribe. More entertaining content such as comedy is likely to be actively played, whereas informative content such as news or documentary may appeal to a smaller audience. Boring, irrelevant, or slow audio such as meeting recordings are unlikely to be compelling sources of content for the game. In other words, even though an arbitrary piece of audio could theoretically be used to create a puzzle for the game, in practice, some editorial decision needs to be made by a human to, at the very least, select appealing content. Furthermore, in order to keep the duration of the game reasonable, the application is best suited for short pieces of audio from sources such as YouTube videos or short news pieces.

One of our participants mentioned that they liked being able to learn from the content of the game as they played. This suggests an interesting alternative application for Audio Puzzler associated with players learning from the content or at least raising awareness for issues expressed in the content in the course of play. While a separate experiment would need to be run to test any learning effect, such a use could be of great interest in, for instance, news content. Already authored news stories, in the form of audio and video packages could be played as puzzles instead. Such an application might have the added benefit of engaging people who might not otherwise have been exposed to the information.

### **Scalability**

One caveat to the technology is its scalability. Audio Puzzler is a labor intensive method for acquiring transcripts insofar as it requires a fair amount of human game playing time to make it work effectively. According to our data, each word transcribed took 4.95 seconds of game play. A related project, the reCAPTCHA system, uses human computation as a security mechanism to transcribe words from aging scanned documents that couldn't be recognized with Optical Character Recognition (OCR). It achieves an efficiency of just 13.06 seconds per word. The reCAPTCHA system has however been shown to scale up to 4 million words per day, with well over 40,000 websites using it. Part of the reason reCAPTCHA scales well is because the overall amount of effort required in one session (13.06 s) is significantly less than for something like Audio Puzzler, where solving a puzzle from a 30s long clip might take the player 5 – 6 minutes. reCAPTCHA also has a different incentive

structure in that it provides benefits to users, like access to some resource such as email, which are different from motivations one would have for playing a game.

With completion rates of 15% in Audio Puzzler, even 1000 players per day would result in only 2.5 hours of transcribed video, and this takes no account of the redundancy needed to properly aggregate a high quality transcript. In practice, 1000 players per day would result in only 30 minutes of high quality transcription per day. Since the video transcription industry also competes on turn-around time and volume discounts this becomes a significant barrier to Audio Puzzler competing with other existing methods. One possibility for scaling Audio Puzzler up would be to use a technique similar to reCAPTCHA and reduce the amount of effort needed per person. This would require a very accurate word segmentation algorithm as well as a completely different game mechanic.

### **Demographics and Game Play**

When Audio Puzzler was first put online it garnered an audience mostly from people interested in novel games. However, as time went on much of the sustaining traffic for the game came from English as a Second Language (ESL) learning blogs which had posted the game as a way to practice English listening skills. While we did look at English proficiency in the laboratory study and found a correlation to transcript accuracy we did not collect this data for the online version and so are unable to comment on how this will impact the transcription accuracy in conjunction with aggregation.

For the online version of Audio Puzzler, the completion rate, that is the number of the times a puzzle was completed divided by the number of times it was

started, was measured at 14.6%. Though we have no baseline to compare to, a completion rate of about 1 in 7 seems somewhat low to us indicating that the game is perhaps too hard still. In one variant we offered to pay people \$0.50 to play the game (and finish it) on Amazon Mechanical Turk (AMT). AMT is a marketplace where microtasks are completed by humans for small amounts of money. The completion rate in this case rose to 23.1%, indicating the monetary incentive was an effective way to motivate puzzle completion, however this came with a slight reduction in overall accuracy. The Medical Error clip transcription data was collected in this way, resulting in a WER of 2.21%.

## **CHAPTER 4**

### **AN ANNOTATION MODEL FOR MAKING SENSE OF INFORMATION QUALITY IN ONLINE VIDEO**

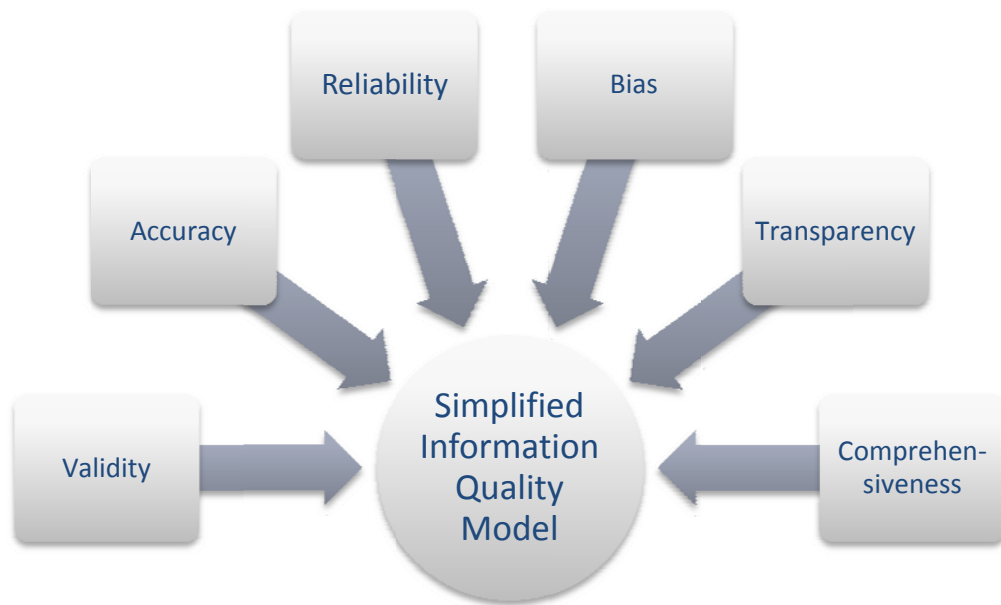
The aim of this chapter is to explicate the conceptual and practical issues associated with the development of the information quality annotation model reified in Videolyzer [23]. A number of challenges are associated with the design of an annotation system related to information quality including how to define information quality and how to operationalize it as a set of annotations that are available to structure people's analyses. And while information quality has been studied extensively in many domains, simplifying its presentation and making it flexible in the user interface is both a hard problem and essential for adoption. In this chapter we present a simplified model hybridized from many others and an operationalization of information quality as a set of annotations ultimately implemented in Videolyzer.

We have synthesized information quality descriptions and ontologies from journalism and communications [44, 60, 65, 80], information science [41, 94, 110, 114], social science [130], health information systems [13, 55], and semiotics [50, 129]. We subsequently pruned our description of information quality to make it less complex and more tractable as an interface aimed at political bloggers and journalists, while still covering as much ground as possible. Our ontology is focused predominantly on issues espoused in journalism and information science and is meant only as a starting point for the analysis and sharing of quality judgments, which we allow always to be augmented by free text descriptions. In the following

sections we will (1) define some of the key facets of the ontology and (2) discuss how these facets are operationalized through annotation. Our goal is not to comprehensively define information quality (a classic problem) but to come to a practical understanding of it as it relates to sense-making and analysis of online video.

### **Information Quality and Its Constituents**

The term *information quality* suggests a degree of excellence in communicating knowledge or intelligence. This degree of excellence can be analyzed along different dimensions identified in information science such as *accuracy, reliability, validity, comprehensiveness, and currency* [114]. In news production, where the conveyance of quality information is paramount, principles such as *accuracy, thoroughness, transparency, and an awareness of bias* guide journalists in communicating quality information [43, 60]. Journalism pursues a practical or functional form of truth which strives to provide the reader with enough information to make his or her own decisions [60]. A difficulty in talking about information quality is the varied terminology used in different disciplines for referring to similar concepts; here we blend nomenclature from both information science and communications / journalism. The key facets of the simplified model are shown in Figure 4.1.



**Figure 4.1. Key Information Quality Facets**

## **Validity**

Validity refers to the (1) well-groundedness and justifiability, or (2) logical correctness of a claim. To a large extent validity has to do with the degree to which the interpretations and inferences about information are sound. According to the two basic definitions, validity can be framed as either a subjective or an objective enterprise. The objective portion of validity is in the logical correctness of inferences; there is an objective and correct way to make inferences from data. The subjective component of validity entails the well-groundedness, justifiability, and acceptability of assumptions and the relevancy of facts to a claim, which may vary from person to person. Even in science, the validity of the conclusions drawn in a particular paradigm of operation are very much a function of the accepted assumptions and methodologies in that paradigm [62].

In journalism validity is best described by the process of verification that journalists pursue as part of their endeavor to provide quality information. By

verifying facts (data) and inferences and by grounding their reporting in these facts, journalists can achieve some measure of validity. Journalists and social scientists practically achieve validity using techniques such as triangulation and corroboration of information sources or by developing a chain of evidence to support inferences [60, 130].

### **Accuracy**

Accuracy is defined in the dictionary as “freedom from mistake or error (correctness)” or as “conformity to truth or to a standard or model (exactness)” [2]. In news information this corresponds to, for instance, correct spelling and the reproduction of quotes in their original form or at least in a form which preserves their essential meaning without de-contextualization [38]. In terms of reporting on an event where the journalist does not have first-hand experience, accuracy dictates that facts or the version of events presented have been corroborated by multiple reliable sources. This corresponds to the second dictionary definition in that the event should be conveyed in a way that is faithful or conforming to the truth of the event as much as that truth is knowable at the time of production. Some discussions of accuracy argue that freedom from mistake or errors in transfer of information is not enough because contextualization of facts is essential to understanding the truthfulness of an event [61].

### **Reliability**

Reliability can be defined as dependability or consistency. For practical purposes the annotation model we develop focuses on the dependability definition

since consistency requires analysis over a longer time frame. Credibility is closely linked to reliability through the notion of trustworthiness (a synonym for dependability), and is a perceived, subjective quality of an information source or object. Trustworthiness is a measure of the truthfulness, fairness, and lack of bias of a source or communication. It can be thought of as a cue for engendering trust, or inducing confidence in or reliance on a piece of information. The other component of credibility is expertise, which is defined as the perceived knowledge, skill, and experience of the source [37, 114].

### **Comprehensiveness**

Comprehensiveness refers to the “completeness of coverage” of a particular subject or issue [114]. In journalism this is analogous to the principle of thoroughness, which involves learning and gathering as much as possible about a topic before sitting down to produce a story [43]. Thoroughness speaks to the degree of comprehensiveness of research on the side of the information producer, which allows for informed information selections. Comprehensiveness also relates to the journalistic principles of fairness and balance which call for the inclusion of different and diverse sources, perspectives, and opinions [60]. Fairness and balance however are not meant to be blindly applied, but to act as guidelines for an information producer to be comprehensive in a way that is faithful to the underlying truth of the story.

The opposite of comprehensiveness is selectivity, which can also be a virtuous characteristic of information depending on users’ consumption context and needs. Selectivity considers the appropriateness and merit of pieces of information

for a particular audience during media production, resulting in a tailored proportionality of information. Due to resource constraints such as time, space, or attention it is also sometimes necessary to pre-filter or select content before publishing it [6, 61]. Depending on the quality and intentions of the pre-filter (the journalist or other information producer), this selectivity may variably affect the context, understanding, and resulting bias of the information.

### **Transparency and Bias**

Transparency and bias are concepts that arise as key components of the journalism of verification [60]. Transparency involves disclosing and describing how information was gathered including sources and how they know what they know, what biases sources or the journalist might have, how decisions were made to include or exclude information from a story, and whether or not there are conflicting accounts of an event [60, 95]. Transparency also includes the notion of acknowledging when information is uncertain or unavailable and what assumptions have been made which may have affected interpretation.

Bias implies a certain “inclination, predisposition, leaning or bent” and can be both individual (e.g. stereotyping) or structural, systemic, and unconscious [38]. Bias has a strong relationship to the sources and other aspects of information that are selected or excluded, but can also include aspects of the communicator’s relationship to institutions and sponsors. Information producers also select information to account for resource constraints in the consumption medium and in the process may introduce bias through the omission or over / under emphasis of information (intentional or not).

## Currency

Currency in information quality quite literally refers to the degree to which the information being communicated is recent and up-to-date with what is known at the time. The most pertinent annotations relating to currency are the timestamp for when the information was originally produced, and whether or not there is newer more relevant information available. For the sake of simplification currency has been elided from the model we have developed for Videolyzer. In other work we have done on information quality in news it has however played a more prominent role [108].

## Practical Design Goals

Taking a cue from the practicality of journalism, the design of the annotation system strives to reduce the complexity of the information quality model for users [129]. The aim is to operationalize analysis of *most* of the aspects of information quality in as *simple* an interface as possible. The functional goal of the annotation system is to provide a way for people to analyze, collect, share, and respond to criticisms of information quality. This involves support for arriving at a consensus about the quality of a source as much as it does for providing a description of the strife or contentiousness of a claim. A central design goal of the annotation system was to maintain expressivity for the user's communication goals and to remain flexible enough to represent most information quality issues without being overly restrictive or abstract while simultaneously leveraging structure from the ontology to provide additional organization and sense-making support. As suggested by Shipman in [100] we took measures in our design to base our formalism on a model

derived from studies of the domain familiar to our intended users (journalism) [23], as well as having non-formal ways to use the system while still providing additional benefit if the formalism was used.

Annotation systems can be characterized by the degree to which the annotation vocabulary is restricted; that is how flexible and open they are. At one end of the spectrum there are completely uncontrolled annotation systems such as collaborative tagging [72, 78] and at the other end are tightly controlled vocabularies used by information architects to tightly organize large quantities of media and information [85]. The advantages of using an uncontrolled approach are that no knowledge of a special vocabulary is needed, people can describe content in whatever way is most meaningful to them in their context and for their use, and there is a low human cost for making a contribution. On the other hand, controlled vocabulary annotation facilitates structured organization in the interface and more formal processing capabilities based on the typed relationships in the ontology [9]. Using controlled vocabulary also avoids some of the difficult issues surrounding inferring meaning from free text comments [12], however this comes at the cost of a greater learning curve and less flexibility for adapting to different individuals' contexts or uses of language.

In order to get the best of both worlds our model supports both unconstrained text comments as well as typed annotations regarding information quality. Typed annotations can always be accompanied by free text reasons or rationale to explain the formal rating. Allowing for free text rationale in the model is important since there are several subjective ratings for which the rationale aid in

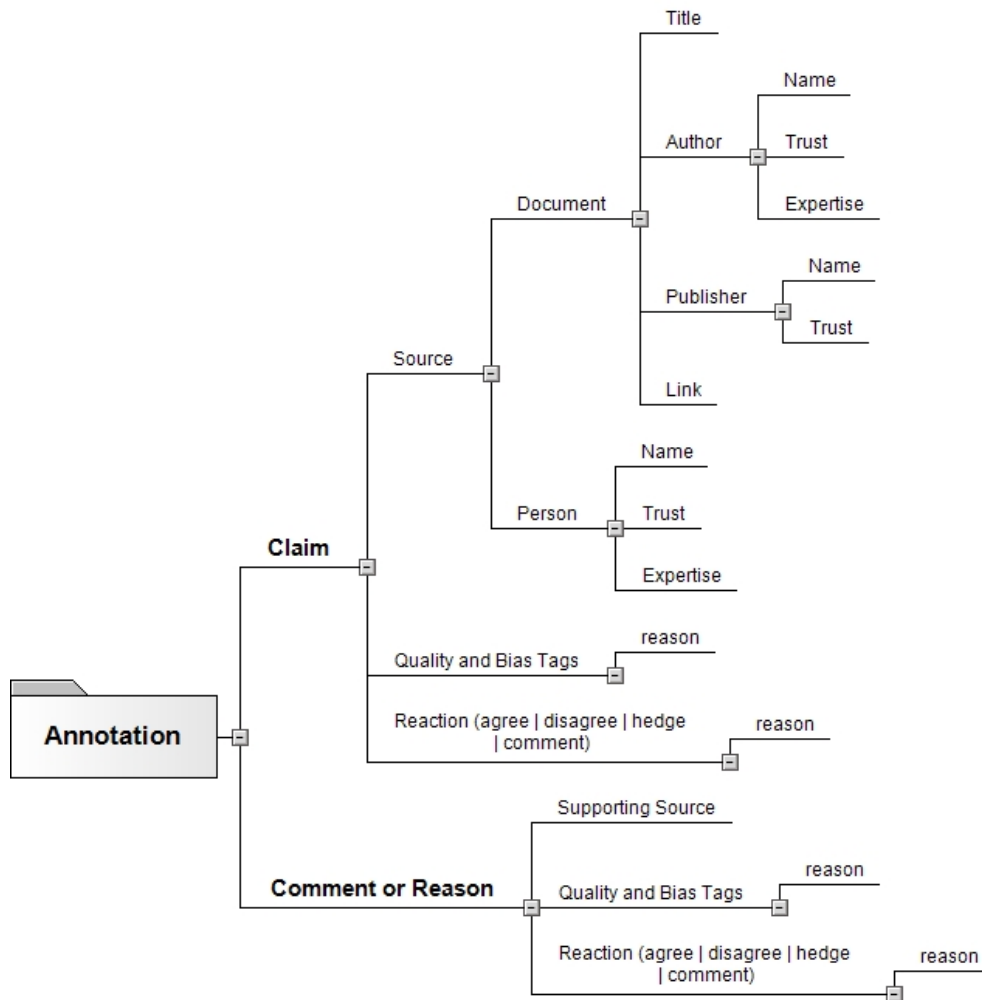
understanding. Also, research in psychology indicates that overconfidence bias in ratings can be mitigated when people are asked to write down their rationale [32].

Issues of literacy explored in related research suggest that by focusing the analytic activity around some key constructs from information quality which are prominent in the interface we can *guide* users towards a better analysis of information quality [99]. In the process, the value of the annotations, both to aid in sense-making and as computable resources, will be enhanced from the structure of the ontology.

### **Information Quality Operationalization**

Here we operationalize the different facets of information quality into the annotation system that is applied to Videolyzer. We have designed the annotations to address the key facets of information quality outlined in the preceding section in as simple a way as possible.

The two high level annotations possible are *claim* and *comment*. Claims represent statements, inferences, or interpretations made in the video and comments are free text remarks made by users of the system. Either can optionally have a *source* associated with them and those sources can be rated according to their *trustworthiness* and *expertise*. Sources are origins of information, whether from documents or people, and can also be described according to relevant elements of the Dublin Core metadata standard [41], such as title, author, and publisher. The relationship between annotations is shown in Figure 4.2. The initial choice between claim or comment minimizes the complexity of the ontology at the interface level.



**Figure 4.2. Relationships Between Types of Annotations**

The rest of the elements of information quality are operationalized as a set of tags divided into two groups: *quality tags*, which are presented as semantic differentials, and *bias tags*. Each tag can have a *reason* for associating that tag. Reasons in turn can have optional supporting sources attached to them. Quality tags (e.g. invalid | valid) are presented as a 3 point scale, the midpoint representing uncertainty. Users may also specify free text tags to add concepts outside the

ontology, however this comes with their lack of semantic interpretability in relation to the other annotations. Tags can be added to comments, claims, and reasons. The various types of tags are shown in Table 4.5. Annotations can be responded to by users in terms of their reactivity (agree, disagree, hedge, comment) as well as by adding quality and bias tags. Thus people interact and negotiate with others through rationalized reactions and by adding tags to other comments and reasons.

**Table 4.5. Tags for Describing Information Quality and Bias**

| Quality Tags               | Bias Tags               |
|----------------------------|-------------------------|
| Invalid   Valid            | Political               |
| Inaccurate   Accurate      | Omitted Information     |
| Bad Premise   OK Premise   | Under Emphasis          |
| Opinion   Fact             | Over Emphasis           |
| Irrelevant   Relevant      | Stereotyping            |
| Unfair   Fair              | Sponsorship             |
| Missing Context   Thorough | < <i>user defined</i> > |
| Unclear   Clear            |                         |
| < <i>user defined</i> >    |                         |

Assessment and annotation of the validity of information hinges largely on the potential for annotators to provide *falsifying evidence* to refute the justifiability or logical correctness of an interpretation, inference, or claim. Within this model, annotators are able to attack the validity of information by criticizing claims, assumptions or premises of claims, and the relevancy and adequacy of facts or opinions being presented as evidence for claims. All of these things are supported through the ability to specify tags, reasons, and supporting sources.

While the validity, accuracy, reliability, and bias of claims and sources are all *directly* supported through annotation mechanisms, issues of comprehensiveness / thoroughness, and transparency are best addressed through the collaborative

interplay and collection of annotations by many people. Given enough annotators with diverse backgrounds, we expect the aggregation of their reasons and ratings will add greatly to the comprehensiveness, fairness, and balance of the information. Moreover, even though the original sources behind a video may not be known, we provide for transparency in a different way. Through the collaborative evaluation of claims and their sources, the transparency of sources, their expertise, and their trustworthiness can be elucidated. Triangulation, corroboration, and chains of evidence can arise through the comparison of diverse sources and reasons provided by different people thus providing a check on the accuracy of information. As individuals agree or disagree with others' ratings and reasons a rich and lasting description of discourse will have cues to all of the elements of information quality in our model: validity, accuracy, reliability, comprehensiveness, transparency, and bias.

## CHAPTER 5

### VIDEOLYZER: ANNOTATION OF INFORMATION QUALITY IN ONLINE VIDEO FOR BLOGGERS AND JOURNALISTS

In this chapter we begin to address the larger issue of information quality by building and studying a video annotation tool, Videolyzer, specifically designed to aid *political bloggers and journalists* in the activity of *watchdog journalism*, the process of combing through and evaluating the veracity of claims in the media. Videolyzer allows users to collectively analyze the quality of online political video and then aggregate and share these analyses with others. Bloggers and journalists can assess aspects of quality in the video, its transcript and annotations according to the model of information quality described in Chapter 4. The advantage of focusing on these users and with political content is that bloggers and journalists are already highly motivated to critique online videos, and political content is rife with claims and issues that would benefit from careful scrutiny and additional contextualization.

We have developed Videolyzer as a visual and discourse analytic sensemaking tool that innovates on many of the annotation issues exposed in the related literature described in Chapter 2 including:

(1) *granularity*, which considers the level of detail of ratings and how different pieces or segments of video can be rated separately;

(2) *multimedia*, by tightly integrating interactions between the video, its transcript, and annotations via the high precision per word time-stamp information using the method described in Chapter 3;

(3) *argumentation*, by integrating a sophisticated information quality semi-formalism (described in Chapter 4) designed to lend structure and organization to the information space while helping users discuss and engage issues of quality collaboratively, and;

(4) *automation*, by incorporating elements of automatic content analysis to help draw attention to areas of the video and transcript potentially worthy of further human analysis.

In the remainder of this chapter we describe additional related work as well as the design and an initial evaluation of Videolyzer looking at the properties of the discourse generated with the tool and the behaviors and perceptions of users regarding the novel features we provide.

### **Related Work**

There is a wealth of related research on video annotation, visual analytics, argumentation support, and the semantic web that has informed the design and development of Videolyzer. As was described in Chapter 2, much of the research on video annotation systems has largely focused on automatic or semi-automatic tagging or commenting on videos without an element of *argumentation support* or *collaborative sensemaking*. Videolyzer extends and synthesizes past work across many domains to marry video annotation interfaces with a notion of helping the user discuss and understand the quality of the information in the video using the semi-structured annotation system described in Chapter 4. Here we describe work relating to argumentation support and collaborative sensemaking which was not covered in Chapter 2.

## Argumentation Support

Recently there have been several video applications in the commercial space such as Viddler and Asterpix, which address the issue of granularity in annotation (i.e. they allow for comments to be anchored to segments of the video). One feature that these systems lack, however, is the ability to combine granularity of annotation with discussion beyond the level of “neat” or “that’s cool.” In some sense these systems are designed to be widely applicable to *any* video content, whereas Videolyzer is targeted toward *informational* content with a concomitant need to understand that information’s quality through an argumentation support system.

Argumentation support and visualization tools are in and of themselves a rich area of prior work (see [56] for a review). Compendium [1] is a widely used argumentation support system which builds off of ideas stemming from the Issue Based Information System (IBIS) methodology of Horst Rittel. The dialogue mapping and conversation modeling approach in Compendium has seen a wide range of applications including design, collaborative meetings, and more recently integration with video meetings in the Memetic system [103]. Memetic combines support for identifying and structuring video conferences based on issues, questions, and arguments / dialogue in the meetings. The highly structured argumentation functionality is integrated with a multileveled video timeline allowing for semantically enhanced navigation. Users can navigate the video based on the argument structure and vice versa.

Trellis is an argumentation system which combines free text descriptions with formal connections between statements [41]. Early versions of Trellis focused

on comprehensively allowing for information source analysis in conjunction with connectives between statements drawn from those sources. Sources could be annotated with their credibility and reliability allowing the system to make inferences about statement support. One of the significant points levied against fully featured argumentation support systems such as Compendium and Trellis is the cognitive effort involved with manipulating a rich ontology [100]. Follow up work on Trellis acknowledged the substantial user effort required to use a formal reasoning representation and explored a range of tradeoffs in designing semi-formal representations which mitigate user effort but which still allow for some machine processing using the structure [12].

### **Collaboration: Social Data Sensemaking**

Collaboration is a significant component of many argumentation support systems. In the POLESTAR system [91], which was designed for intelligence analysts, collaborative review of intelligence reports was facilitated through a semi structured mechanism for feedback including agreement / rebuttal, request for clarification, and request for supporting evidence. Videolyzer is similar to POLESTAR in its effort to enhance the quality and validity of inferences through peer-review.

Collaborative annotation has also been employed in helping to make sense of a variety of information visualizations. This has been termed *social* data analysis, and typically involves many people working asynchronously to find and comment on interesting features (e.g. trends, outliers, comparisons) in visualizations. For

instance, the Many Eyes visualization system [117] was explicitly developed as a forum for people to discover and discuss interesting features of visual data sets.

Another system oriented toward asynchronous discussion and annotation of information visualizations is sense.us [49]. Most interesting in sense.us was their notion of “embedded discussion” which anchors sub-discussions to particular view states of the visualization. We also utilize this design objective in Videolyzer to help ensure users have focused discussion on different parts of the video and transcript.

## **Videolyzer System**

### **Annotation and Interaction model**

The interface of Videolyzer translates the information quality model described in Chapter 4 into a useful set of actions that users can take to interact with information in the video and in other annotations in a semi-structured way. The design goal of these interactions is to allow for an overall enhancement of the understanding of information quality (1) by allowing users to provide corroborating or refuting evidence for claims and comments thereby giving rise to triangulation or chains of evidence, (2) by facilitating fact checking and accuracy evaluation using alternate sources, and (3) by allowing for the expansion of different perspectives (i.e. comprehensiveness, fairness, and balance), and sources (transparency) around an issue.



Figure 5.1. The tagging panel UI with quality, bias, and collaboration tags.

The primary elements of the ontology exposed in the interface are *claims*, *comments*, *tags*, *reactions*, *reasons*, and *sources*. The two annotations initially available in the interface are claims and comments. After these are added, the other types of annotations become possible depending on ownership (see Table 5.6). Also, in addition to the tags described in Chapter 4, *collaboration* tags have also been added (Figure 5.1) so that users can communicate with each other about whether an annotation needs a source or a review.

Table 5.6. Annotation types and associated actions depending on annotation ownership.

| Annotation Type | Actions on Own Annotations                            | Actions on Others' Annotations             |
|-----------------|---|--|
| Video Claim     | Add Source, Add Tags, Add Reaction                    |  |
| Comment         | Add Source, Edit Text, Delete                         | Add Tags, Add Reaction                     |
| Tag             | Add Reason, Delete                                    | Add Reaction                               |
| Reaction        | Add Reason, Delete                                    |  |
| Reason          | Add Source, Edit Text, Delete                         | Add Reaction, Add Tags                     |
| Source          | Specify type, Edit Title, link, Rate expertise, trust | Edit Title and link, Rate expertise, trust |

## Automatic Content Analysis

Videolyzer is primarily focused on helping users do manual information quality evaluations of video. Automatic methods of information quality annotation are largely impractical because many aspects of information quality have a degree of subjectivity, context, and connotation in their interpretation and evaluation. However we are interested in investigating the notion of automated suggestive annotation oriented towards concepts of interest to media studies. Furthermore, automatically extracted structure from video has been shown to aid navigation and information retrieval scenarios, which we would expect to enhance the usability of our application [47]. As such we have incorporated some automatic methods in order to examine if they can help *guide* human attention to areas of the video or transcript that may deserve more critical attention. We use content analysis algorithms to detect faces in the video, as well as references to people, money, comparisons, and modal verbs in the transcript. These are rendered as colored markers over the transcript (Figure 5.3) and timeline (Figure 5.4).

We felt face detection could be important because many sources of information in video are people and a face can be a proxy for helping the annotator find these sources. We use the OpenCV face detector to preprocess the video. It is run on each frame separately and the temporal stream of detections is median filtered to remove any spurious detections.

We also chose to extract references to people, money, comparisons, and modal verbs from the transcript because (1) media studies suggests that these features may lead to interesting analytic insights about the quality of information

presented [38, 52] and (2) it is possible to automatically extract these features with a fairly high degree of reliability. References to people and money are extracted using the Open Calais web API from Thomson Reuters [3].

Comparison and superlative words in the transcript can act as further cues to a human rater interested in suspect statements. For instance, dangling comparatives such as “larger” or “better” beg the question of what is being compared to; there is no indication of larger *than* or better *than* [52]. Similarly, superlatives such as “most” or “best” can be vague if they are not used in reference to some measure. Comparisons in the text are detected by running a Part Of Speech (POS) tagger on the transcript and then noting words detected as either comparisons or superlatives [8].

Text modality, which is expressed through modal verbs (e.g. can, ought to, could, shall, may, should, might, will, must, would) together with their negations and adverbs is an indicator of the degree to which a speaker or communicator is committed to their claim [38]. The choice of modal verbs can indicate the attitudes, judgments, and political beliefs of a writer or speaker. Verb modals can vary from weak to strong (i.e. possibility to necessity). For instance, a weak claim might involve the phrase “ought to” whereas a strong claim would use the verb “must.” We developed a routine to detect modal verbs and their strength by filtering the textual transcript against a list of modal verbs identified by linguists [66].

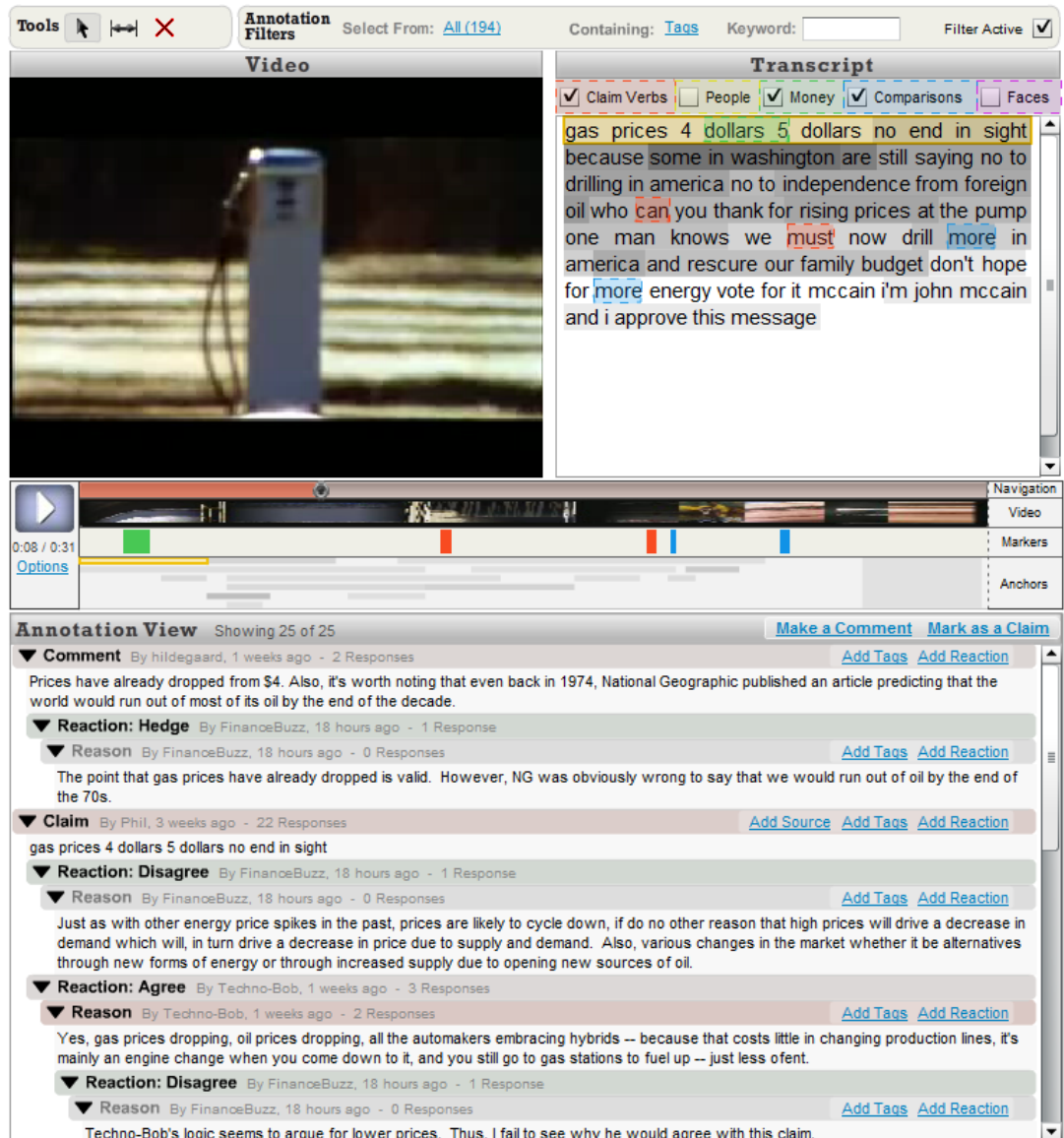


Figure 5.2. Overview of the Videolyzer interface showing (clockwise from upper left) the tool bar, filter options, transcript, timeline, annotation view, and video playback area.

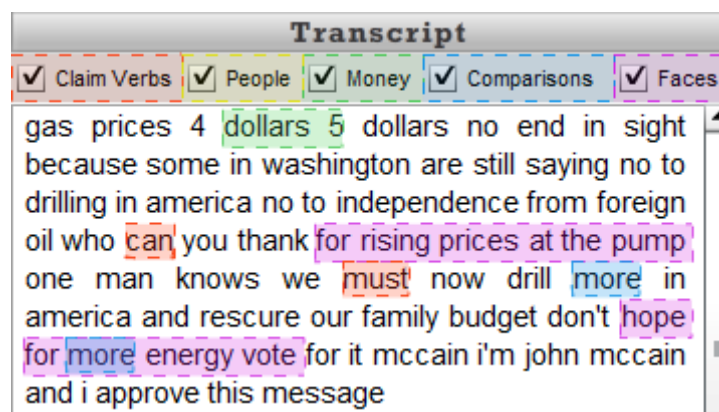
## Videolyzer Interface

The videolyzer interface is designed to tightly couple the interactions occurring across three multimedia indices: the transcript, timeline, and annotation views (Figure 5.2).

### Transcript View

A precisely time-stamped transcript is included in the Videolyzer interface to facilitate fine-grained commenting (Figure 5.2 and Figure 5.3). The video transcript is gathered using Audio Puzzler, described in Chapter 3 [28]. Automatic methods for acquiring the transcripts could also have been used although usability and utility become problems for automatic transcripts with lower accuracy [87].

Time-stamp information allows for a very tight coupling of interactions between the transcript and the timeline. Actions such as hovering and selection are mirrored on the other view, which we expect to aid navigation and anchor creation on the video. For instance, users can select a piece of text as they normally would on the transcript and an anchor is added to the transcript view as well as the timeline view (Figure 5.4). Anchors are where further annotations are attached to the transcript or timeline. They are rendered over the transcript as gray boxes, the darkness of which is modulated by the number of annotations on the anchor normalized by the maximum number of annotations on any anchor for the video. These anchors are also semi-transparent so that areas of the transcript with more annotations appear darker. This is a form of social navigation allowing users to



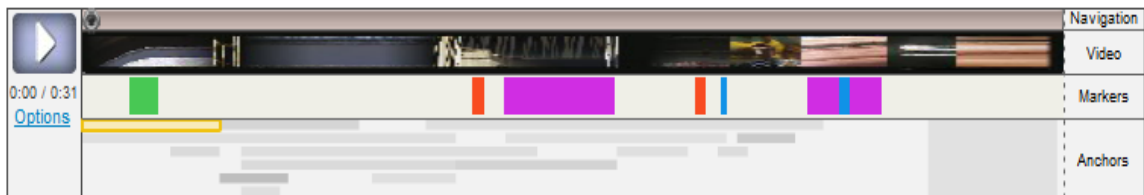
**Figure 5.3. Color coded, automatically detected markers shown on the transcript view.**

orient and navigate to anchors containing more annotations [29].

### Timeline View

Navigation through the video is accomplished using standard video controls such as play and pause as well as by using a stacked timeline representation which shows four types of information (Figure 5.4). The top level of the timeline, labelled *Navigation*, is a standard video thumb which allows for non-linear seeking within the video. The second level down is a stripe image indicating the pixel colors of the video at that point in time. The space efficient stripe image clearly shows boundaries between shots and is meant as an information scent to aid the user in navigating between segments of the video. The *Markers* level shows the same automatically extracted markers as are shown over the transcript (Figure 5.3) but synced to the time of the video. The bottom most level of the timeline shows all of the anchors that have been added to the video. Anchors on the timeline are rendered with the same social navigation information as those shown over the transcript. Anchors can be rendered as overlapping, semi-transparent areas, or as a “stacked” view. The stacked view algorithmically determines how many sub-levels of the timeline are needed to avoid

The bottom most level of the timeline shows all of the anchors that have been added to the video. Anchors on the timeline are rendered with the same social navigation information as those shown over the transcript. Anchors can be rendered as overlapping, semi-transparent areas, or as a “stacked” view. The stacked view algorithmically determines how many sub-levels of the timeline are needed to avoid



**Figure 5.4. Stacked timeline view showing levels for navigation, stripe image, automatically extracted markers, and anchors.**

overlaps between anchors and then sub divides the height of the anchors level accordingly.

Hovering over the timeline brushes the other views. It shows the frame at that point in time in the video player area of the interface (Figure 5.2 upper left), highlights the word using a transparent gray box in the transcript area at that time and shows a tracking popup above the timeline with the three words in the temporal neighborhood of that frame.

### Annotation View

The annotation view (Figure 5.5) shows the set of filtered annotations attached to the currently selected anchor on the timeline or transcript. If no anchor is selected, the annotation view shows the global annotations which apply to the entire video. The view is structured as a set of hierarchically composed annotations which can be expanded or contracted by clicking the arrow at the left hand side of

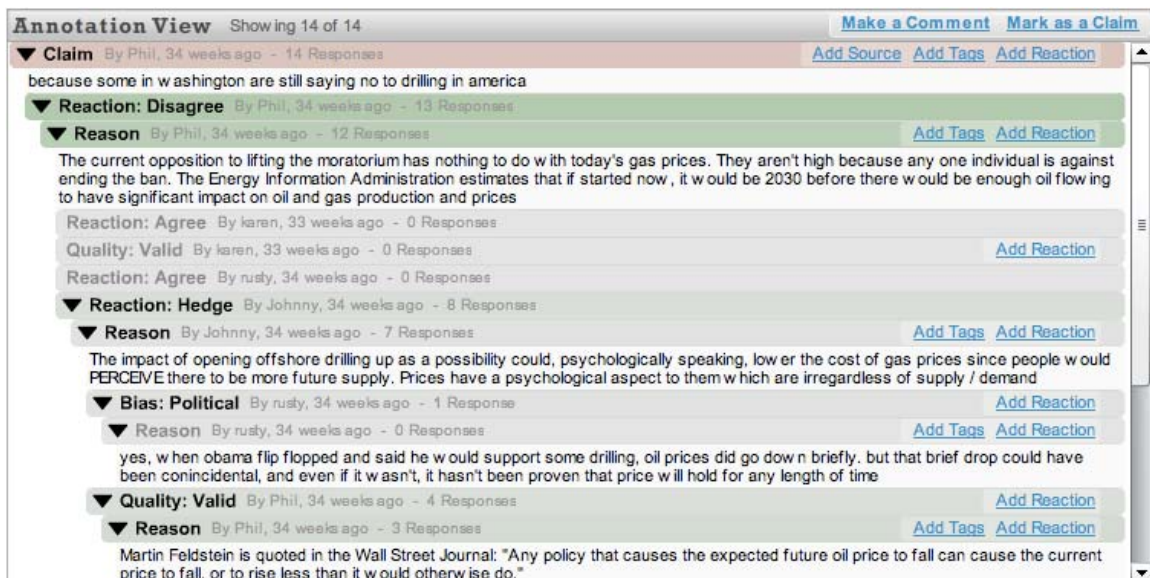


Figure 5.5. The Annotation View showing a claim that has lost support based on the series of annotations responding to it.

each annotation. Each annotation shows its type, the username of whoever added the annotation, how long ago the annotation was made, and the number of responses to that annotation.

At the right hand side of each annotation is a list of action buttons for interacting with the information. As described earlier, the actions available depend on the type of annotation and whether or not the user has added the annotation or someone else has added it. The view is designed this way to minimize memory and cognitive issues associated with learning the ontology. All initial actions are visible in the interface, however some of them expand. For instance, if the user clicks “add reaction” a drop down will show the options of “agree”, “disagree” etc. Also, clicking “add tags” pops up the tagging panel shown in Figure 5.1. At the top of the annotation view are the two primary actions available when any anchor is created or selected “make a comment” and “mark as a claim.”

The coloring of each annotation indicates the degree of support of that particular annotation in the system with shades of green representing more support and shades of red indicating less support. For example, in Figure 5.5 there is a claim that has been identified at the top which is colored slightly red because someone reacted in disagreement to the claim and the reason for this disagreement was corroborated by other users (i.e. it in turn is slightly green).

Degree of support values in the annotation view are computed bottom-up from the annotation hierarchy. Each type of annotation is assigned an initial quality weight (QW) in  $[-1, 1]$ . For instance, negative valence tags such as “invalid” start with a QW of -0.5 whereas positive tags such as “accurate” have a QW of 0.5. For

each node in the hierarchy its quality score (QS) is computed by summing the QWs of its children. The sum is divided by a constant and capped at -1 and 1 so that a child does not overly impact its parent. Annotations need corroboration from other similarly valenced siblings to substantially effect the parent's QS. The QS is then mapped to the color red if less than 0 and to green if greater than 0.

The quality weight of the parent node is also modulated by the normalized child QW sum so that the support that node transmits to its parent in turn takes into account the values of its children. We also consider changes in sign of the parent child relationships. So, for instance, if an initially negative node has accumulated a negative quality score based on its child annotations, the value it transmits (its QW) will be less negative. Intuitively, a negative annotation that many disagree with or have rated down has less negative impact on its own parent.

## **Filtering**

Filtering in Videolyzer is provided as a way to narrow down the visible annotations in the annotation view, timeline, and transcript (Figure 5.2 upper right). When particular annotations are filtered out they are still rendered at reduced saturation and brightness in the annotation view so that users can see the context of the remaining annotations. Anchors with no annotations in the filter are not rendered, thus reducing visual clutter.

The filters were designed to facilitate users finding “interesting” annotations to respond to. The filtering mechanism is a logical AND of three components: “select from”, “tags”, and “keywords.” Specifying tags allows the user to search for

annotations tagged with any of the quality, bias, or collaboration tags detailed earlier. Keyword filtering applies to any annotation with free text to search.

The “Select From” component of the filter allows users to specify the set of annotations being shown. Users can select from “All”, “My”, “Highly Rated”, “Contentious”, and “Feeling Lucky.” To detect highly rated annotations, the system thresholds the QS of any given annotation. “Contentious” annotations are detected when the variance of QWs on an annotation is above a threshold. The “Feeling Lucky” filter randomly selects a set of annotations.

### **User Study**

We conducted a user study to help gain an initial understanding of the user experience of the application, whether it was usable, and if it was useful to political bloggers and journalists. At the interface level specifically, we were interested in looking at the usage and behavior around the transcript, timeline, and automated suggestion markers. And from the content, we wanted to explore the collaborative dynamic of the discourse generated with the tool, including people’s information quality evaluations. We collected data from questionnaires, interaction logs, screen recordings, interviews, and annotation content.

### **Procedure**

The study was conducted in a usability lab on a dual monitor computer so that participants could have one monitor for using the application and one monitor for looking up information in a browser. Each participant first filled out a background questionnaire soliciting information about their experience and

education in journalism, their online news and video consumption and commenting behavior, and their interest in a variety of political issues. Then the participant had ~15 minutes to read through and watch videos from a tutorial explaining all of the features of Videolyzer. The participant was then given a warm-up period where they used the interface to accomplish a list of tasks to ensure their familiarity with the different features.

Finally, the participant was given two videos (one at a time) to analyze with the interface. Their task was not specified precisely, but instructions were given to analyze the quality of the videos and to have a discussion with other users. The experimenter left the room and checked back on the participant in 15-20 minute intervals until the participant felt comfortable that they had “finished” their analysis. Afterwards the participant filled out a questionnaire and answered some questions in an interview. The questionnaire included ratings for various aspects of their experience on a 7 point Likert scale (1 negative and 7 positive). Subjects were run in a sequential order in which each participant could see the annotations of all prior participants. In addition to the initial and final questionnaires and interview we logged all interactions that the user had with the interface including the type and quantity of all annotations made. We also screen recorded sessions to look for interaction patterns and usability issues.

## **Content**

The videos chosen for the study were taken from YouTube and were selected for their multivalent properties. We focused on produced content since it tends to be more information dense. We chose videos which relate to the energy issue in the

United States. Energy policy is a timely and relevant topic of interest to many Americans so we thought participants would already have some familiarity with the topic. Also, energy issues lend themselves to both objective analysis through statistics and subjective analysis via different perspectives on the issue.

The warm-up video was a 30 second advertisement from the American Petroleum Industry. The first video in the study proper was a 30 second political advertisement from the Republican nominee for President, John McCain, called “Pump.” This ad had garnered attention in the media as containing many weak claims so it was a natural candidate for generating discussion between participants. For this video the commentary was seeded with information gleaned from a FactCheck.org<sup>1</sup> expert analysis of the video as well as with comments from three pilot participants. The final video in the study was a 2.5 minute mini-documentary from the Heritage Foundation about U.S. energy policy as it relates to ethanol production. This video was not seeded with any annotations.

## **Participants**

We recruited 10 participants from both local political blogs and news outlets through email and via subsequent snowball sampling. Six had a degree in journalism or had on the job experience in journalism and four did not. All but one participant were male and ages ranged from mid 20s to late 50s. All of our participants could be considered “news junkies” in that they all read online news several times per day

---

<sup>1</sup> FactCheck is a nonprofit advocate “that aims to reduce the level of deception and confusion in U.S. politics.” <http://www.factcheck.org>

and watched online political videos at least 2 times a week (and most watched much more than that). Everyone was a frequent reader of comments on blogs and videos, and all but one person had done extra research to support a comment “often” or “sometimes.” Also 8 of the 10 participants said they had spent extra time to verify facts presented in news and had cited original sources of information in comments “often” or “sometimes.”

## **Results**

### Interface Usage

The Transcript was *strongly* preferred for both creation of new anchors and selection of existing ones. In fact no new anchors were created on the timeline, and 43 were created on the transcripts across all participants and both study videos. The transcript was also used much more heavily for navigation and selection of anchors with 82% (151 / 184) of the selections. Part of the reliance on the transcript could be associated with the fact that the videos used were primarily audio information centric [68]; other types of content, such as very visually oriented video, could elicit different interaction strategies.

We asked participants about their use of the transcript in the interview. They thought it was much easier to mark areas of interest on the transcript, that the anchors were larger and easier to select, and that it was more suited to navigation in general. Users had a bias toward interacting with textual information, so the textualization of the video through the tight coupling of the transcript with the timeline seemed to be of great benefit. Several participants said that it helped them

find interesting areas of the video to comment on: *“For finding what I was interested in commenting on in the video the transcript was best,”* and *“Having the transcript there really allows you to get directly to what they’re saying without having to go in there and listen.”*

Participants thought that having the integrated transcript allowed for an enhanced discussion of the video. One participant indicated it was easier to stay focused on what was actually said and to be commenting on the same thing as others by looking at the highlighted anchor on the transcript. Another liked the ability to highlight specific phrases where the meat of the content was. Not everyone thought the transcript was unequivocally great though. *“Most people watch videos because they don’t want to deal with text,”* remarked one participant. The transcript also lacked punctuation such as capitals, periods, and paragraphs and could be difficult to parse. Sometimes the textualization of the video (including all the annotations) was distracting from the visual of the video.

#### Suggestive Annotation and Use of Automatic Markers

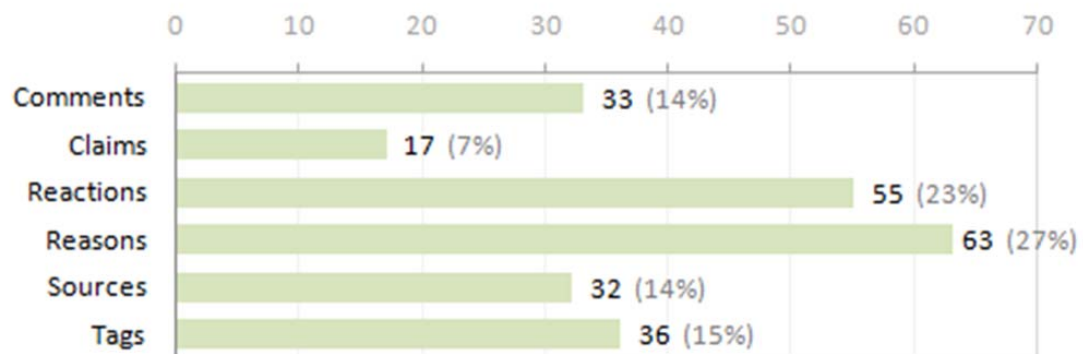
Automatically extracted markers were only used by 4 of the 10 participants. Among the reasons others gave for their disuse were that the videos were too short and that looking at the text was sufficient for finding what they wanted to comment on. One user mentioned that the markers didn’t correspond to things that he cared to comment on. A few others thought that the markers and their colors were overwhelming and were more than was needed to use the system effectively.

Several participants thought however that for longer videos the markers could be helpful for highlighting different types of arguments, or even as simple

visual metrics for comparing videos. One participant acknowledged their potential usefulness saying, “*People equate facts to mentions of money, and feeling to how many times you mention those verbs.*” Indeed, one user told us in the interview that he had used the money marker to hone in on an area of the transcript that he ultimately wanted to comment on. He also tried the claim marker to jog his memory about where they had said something in the video, but ended up deciding not to comment based on it. Two other participants were observed to have used the highlighted markers in the process of navigation to hone in on anchors in a vicinity where they were interested in exploring the discussion.

#### Annotations and Annotation Behaviors

Across all 10 participants there were 236 annotations created including comments, claims, reactions, reasons, sources, and tags. The breakdown and distribution of their usage is shown in Figure 5.6. At the top level, comments were used almost twice as often as claims were marked. At the response level, reactions and reasons were the most heavily used aspects of the ontology representing half of



**Figure 5.6. Annotation Type Usage and Distribution.**

all annotations. Tags were used primarily for criticism (not support) of other annotations including some use of the collaboration tags, “needs review” and “needs source.” Tags were only used by 6 of the 10 participants however.

The ability to source comments and provide reasons was a powerful feature and one used by all participants. In the user experience questionnaire users rated their ability to see and access the sources on which others had based their annotations as 6.60 / 7.0 with a stdev of 0.7. This is perhaps best summarized in one participant’s words, *“I felt like my entry was not complete without one [a source] ... I needed to support my claims with more expert arguments.”* There was also a good amount of corroboration and refutation that occurred between participants using sources. One person used the citation that another had previously added to refute that person’s own comment, an activity that seemed to delight him. Another participant reflected on his use of sources, *“It turned out the biggest thing I was going after, the articles I easily found weren’t supporting the point. It forces you to think about what you’re doing.”* The process of backing up arguments with reasoning and sources made participants more careful and reflective about the information they were using in the system.

One of our concerns with implementing a formal argumentation system was that the expressivity of the user could be compromised [100]. What we found though was that the semi-formalism that we designed was successful in letting people express and discuss the various aspects of information quality. The mean rating for, “I felt I could express myself adequately with this tool.” was a 5.89 / 7.0 with a stdev of 0.93. The formal representations weren’t used by all however. In fact

there were two participants that only used comments and sources, but not reactions, reasons, or tags. One participant recognized that he had agency over the degree of formality of usage, saying, "*If you just want to make a comment, you make a comment, you don't have to put sources. Maybe you just want to get other people talking.*" Some further observations about the use of the ontology are made in the following section.

Granularity and the specificity of annotation was another feature that many users appreciated. Several users mentioned that they liked being able to focus on a particular part of the video or to highlight specific statements and comments. It helped to organize the commentary into more focused and relevant threads. One participant noted, "*You're all launching off of an individual phrase or selection or image and so you're all on the same page.*" Participants thought this helped at least somewhat to keep the discussion on topic.

Asynchronous interactions and collaborations among participants were observed in usage and by looking at the content produced in the annotations. Users were clearly influencing each other and building on each other's arguments. One participant remarked, "*As I come into this I see myself not just as a person that's gonna give my 2 cents, but my 2 cents are going to build up a potentially more interesting argument over time.*" There were instances of joking, emoticons, and socializing and since we recruited from a local pool of participants several participants acknowledged that they recognized others by screen name. They referred to each other by name and one participant admitted that he let his knowledge of these other people influence how he responded to them in the system.

Participants responded to others with questions like, “*Where’s your source?*” and in some cases replied to others’ questions with answers or links to outside sources. Another way that users could interact with each other asynchronously was by reacting to another’s annotation in terms of agreement, disagreement, or hedging. Of the 55 reactions recorded, 41 of those (75%) were reactions to other people’s annotations. The reasons behind these reactions served to request additional information, sources, or parameters, to expand the issue, to add a caveat, to clarify a remark, or to add an alternate explanation. In general, reactions and their reasons expanded the comprehensiveness of the issues being discussed as participants brought individual knowledge, sources, and perspectives into the conversation.

### User Experience

Participants rated their “overall satisfaction with the application and what it can do” as 5.50 / 7.0 with a stdev of 1.18. All other ratings are shown in Figure 7. For the most part users indicated that it was easy to navigate, read comments, and to annotate things. The last participant felt that anchors and annotations (almost 200 annotations at that point) were becoming dense enough to make navigation somewhat harder. A few users felt that the interface was over featured and felt overwhelmed with everything that it could do. The interface is clearly not suited for the average casual user. However, participants also acknowledged that they could fairly easily learn the core features and lexicon of the system within the 25-30 minute training and warm-up period. The rating for “I would be able to produce a detailed analysis of the quality of a video using this tool.” was 5.89 / 7.0 with a

stdev of 0.93, so users had indeed recognized the power of the tool and had gotten comfortable with the interface by the end of the study.

The learnability of the ontology used to drive the semi-formal interactions was somewhat of an issue for some users. Observations of the screen recordings of users showed that 6 of them were seen to hesitate and visibly vacillate their cursor between options in the ontology as they made a decision about what facet was most appropriate. One user was observed to have checked tags for “unclear” and “missing info” but then went back and unchecked them in favor of “invalid.” It’s clear that there is additional cognitive overhead for using the semi-formal structure. At the same time, this structure had clear benefits for some. One participant remarked that, *“It is my reaction that I’m responding to this claim and then I have a reason that I’m providing and then I add the source to the reason. And that’s a nice structure ... I liked that.”*

A key design goal of Videolyzer was to enhance the understanding of the

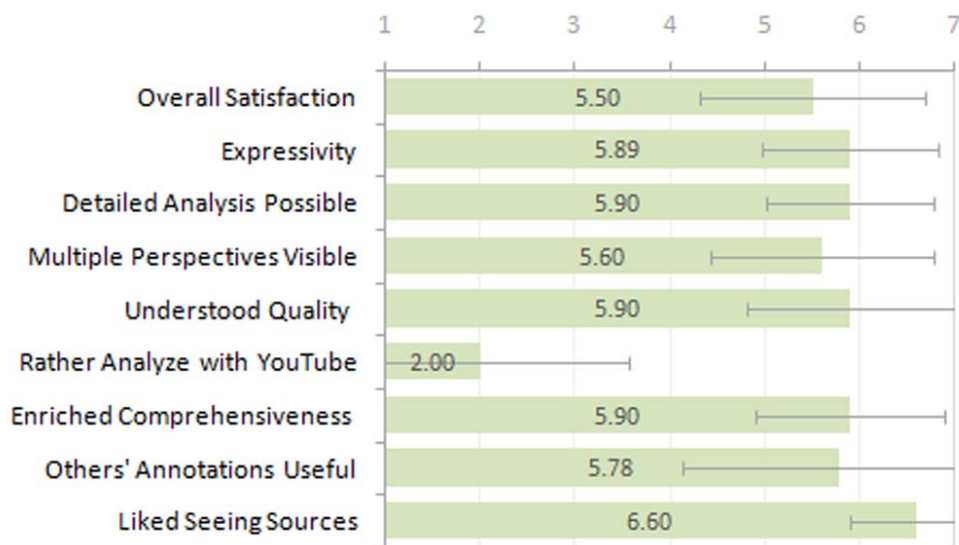


Figure 5.7. User Experience Ratings

quality of the information people are consuming by providing the ability to see comprehensive information as well as diverse and varied perspectives on an issue. Users rated “I could see multiple perspectives on an issue by reading the annotations.” a 5.60 / 7.0 with stdev of 1.17, and “The annotations made the information in the video more comprehensive / complete.” a 5.90 / 7.0 with stdev 0.99. Also they rated “I felt I understood the quality of the information present in the video and its annotations better using this tool than I would have with a standard tool like YouTube.” as a 5.90 / 7.0 with stdev 1.10. Of course we can’t make any claim about an objective change in the information quality of the video, but these ratings indicate that participants’ *perceptions* of the quality of the information in the video were sharpened. Participants’ ability to see multiple and diverse perspectives on complex issues was enhanced. One remarked during the interview that, “*There were some thoughtful comments that made one consider different angles on the material.*”

## **Discussion**

The semi-formalism for quality analysis saw significant usage and did not substantially detract from people’s ability to express their evaluations. There were adequate opportunities for free text explanation including a familiar “comments” mechanism. People could choose to use the elements of the formalism that made most sense to them, could express evaluations of quality through free form text, and were not penalized for not using the formalism.

The ability to source and see sources was by far the most widely liked feature of the system. We suspect this is in large part due to the characteristics of our users;

media savvy bloggers and journalists that know the value of buttressing and guarding arguments with thoroughly researched information. In fact one journalist went so far as to suggest she didn't want to waste time looking at any comments that weren't sourced. She suggested that the application would work best as a shared application in a single newsroom, allowing reporters to leverage each other's sources and expertise to do research and analysis. Being able to add sources gave the interface a "serious" side that appealed to journalism professionals, and seemed to add to their acceptance of the technology. Designers of such interfaces for journalists or serious bloggers may want to consider this and other values central to the practice of journalism in their design [60].

Automatically extracted and highlighted markers such as money or comparisons were of some benefit for users but may have added to an already overwhelming interface. While there were some interesting uses for navigation and comment creation their overall usefulness may be marginal compared to other more important features such as having the transcript tightly integrated with the video. Additional work is needed to assess whether the markers would be helpful for the analysis of longer videos or if there are other markers that are automatically extractable and better correspond to the types of things people are interested in.

### **Limitations**

One limitation of Videolyzer in its current form is that transcripts must be obtained through Audio Puzzler before the video can be used in the application. This limits the ability of a user to use an arbitrary video, and means that either (1) a centralized repository of pre-processed, transcribed videos needs to be assembled

for use or (2) we create a degenerate interface for videos without transcripts and provide a link to the transcription game. If a user is interested enough, they may choose to play the game before looking at the video.

Another limitation is the scalability of our interface representations. We designed our views for use with short (upto ~10 minute) internet videos, but in reality Videolyzer could also be useful for longer documentary type videos. One difficulty here is the scalability of the timeline representation for much longer videos, since for simplicity we currently don't provide for zooming on the timeline.

The annotation view and the way in which hundreds of comments can be aggregated and represented is another difficulty with moving toward an internet scale application. The division of discussion across many anchors is perceived to keep commentary more focused and reduces the number of threads in any one anchor, but there is a limit to this. All tags and reactions (38% of annotations) can be directly aggregated using the ontology, but comments and reasons (41% of annotations) are plain text and need further processing before aggregation. This can be done using content analysis to compute derivative facets such as positive or negative sentiment in the text. In the next Chapter we will show the development and evaluation of alternate views and aggregate representations which reduce complexity further.

## CHAPTER 6

### MODULATING VIDEO CREDIBILITY THROUGH ANNOTATION

#### SYNDICATION AND SIMPLIFIED PRESENTATION

In this chapter we develop and evaluate methods for the *syndication* of the video watchdog information collected using Videolyzer to a less engaged class of users using our system Videolyzer CE (Consumer Edition). Our goal is to enable sharing of the knowledge of interested watchdogs with a wider population of users and in the process modulate perceptions of information quality. We do this by developing simple visuals which indicate aggregated activity levels and polarity of evaluations (i.e. positive / negative) shown in-line with videos as they play. More interested users can interact with and drill into the visualization for the details of the evaluations including tags, sources, and comments. In order to understand the influence of this visualization on casual video consumption we also evaluate its impact on the credibility of the information presented in the video as compared to a control presentation of the video.

#### Related Work

Information quality has been studied in a variety of contexts such as Wikis [111], social media [67], and traditional news stories [64]. While some aspects of information quality are objectively verifiable, others such as credibility (i.e. belief) are *perceived* and *subjective* notions of quality and as such can be modulated on an individual level [34]. Belief in particular can be thought of as a person's estimate of the *subjective probability* or certainty that a proposition is true [126]. The focus of

our work here is the design and evaluation of visual cues which may engender constructive changes to perceptions of belief in informational video (e.g. by cueing people to poor quality information in video). An extensive review of research on attitude and belief change can be found in *The Handbook of Attitudes* [126].

Recent work looking at Wikipedia has suggested that users' perceptions of trustworthiness and credibility of information can be impacted by detecting and then visualizing edit activity and reputation information using relatively simple visual dashboards [58, 92, 112]. Nakamura et al. postulate that a similar effect should hold for social annotation data showing the polarity of time-stamped textual responses to video information, however they fail to present a rigorous evaluation of the claim [88]. Nakamura's implementation also exposes several difficulties when dealing with unstructured video comments and sentiment detection including an inability to discern whether the sentiment of a comment was in response to the original video or to other comments.

Some of these difficulties are avoided with more explicit video evaluation information such as that collected by Videolyzer [27], which includes hierarchically organized quality tags, sourcing, and free text comments. However the visual complexity of Videolyzer and its orientation toward motivated bloggers and journalists means that it is inappropriate for casual users to benefit from its rich annotation information. Here we consider a model where videos would be annotated using a structured tool first, but then these annotations would be syndicated to more casual users via simplified and aggregated representations of the annotations.

## Visualization Design

In the development of our visualization we drew on work in dashboard design [33] and traditional broadcast graphics, which contextualize video information with maps, names, and titles, but for the most part do not provide any notion of social quality evaluation. Our design goal was to distill a detailed hierarchical and collaborative evaluation of quality into a set of simple cues which could be useful to viewers' assessment of a video's quality. Design decisions included both what data to visualize as well as whether that data should be immediately visible or only visible upon engagement and further interaction.

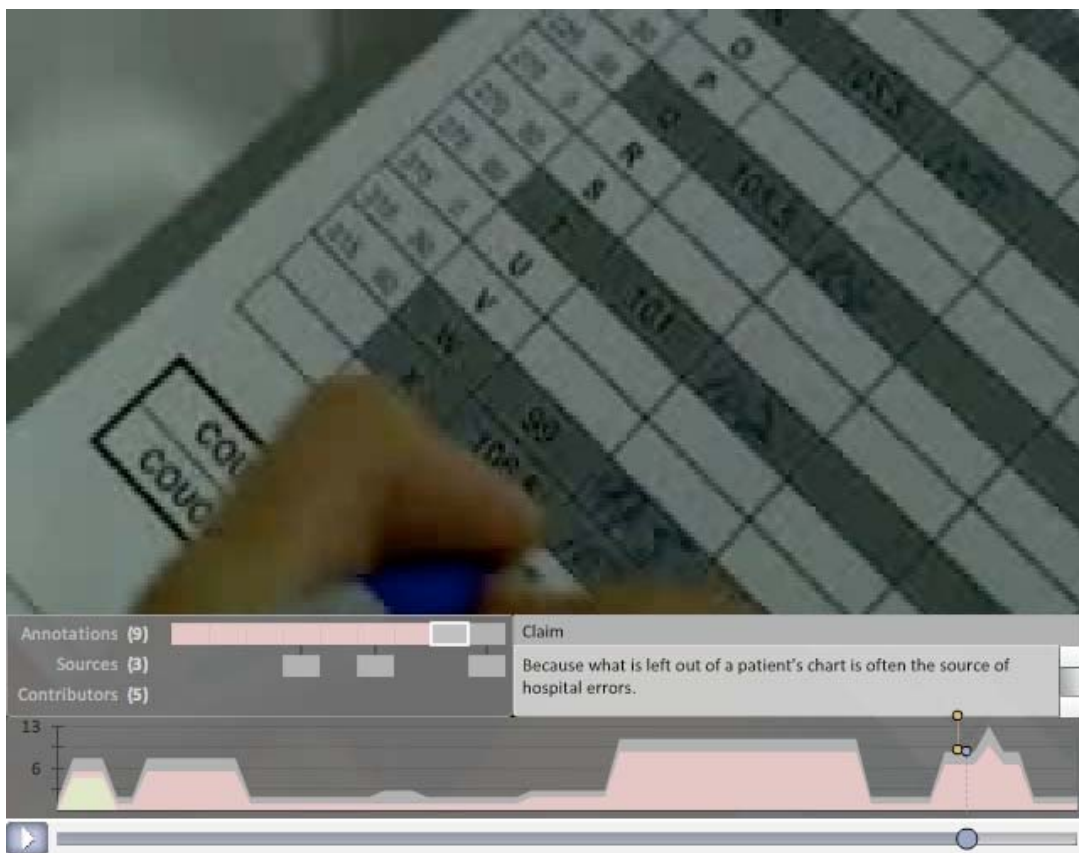


Figure 6.1. Experimental video player showing annotation polarity graph and detail panes.

Prior work on discussion visualization suggests a range of quantifiable metrics for the characterization of the structure and content of online discussions such as *size (i.e. breadth, depth, number of messages and contributors), recency, activity level, anonymity, stability, and tone* [7, 58, 106]. The ability to detect these features automatically rests both on the degree of structure in the commenting system as well as the robustness of content analysis algorithms (e.g. sentiment or affect recognition).

In order to reduce consumption bandwidth as well as to maximize the potential for showing cues relevant to credibility we organized cues into three levels of successive detail. We chose to focus the initial visualization on aggregate measures: *activity level* and *annotation polarity*, with interactions revealing additional information such as *use of sources, number of contributors*, and ultimately individual *quality tags, comments, and sources*.

Activity level, an honest signal of interest, indicates areas of the video that have generated more or less discussion and thus might be worth investigating. Polarity shows whether people have evaluated a section of video as positive or negative. Evidence and sources were included because of their expected impact on credibility evaluations [36, 79]. Finally, the number of contributors was included in order to indicate if the activity or polarity of annotations in one area was the result of one person or a diversity of opinion. Our purpose was not to study the individual effects of these cues but rather to understand if, take together, they can impact the perceived credibility of the video.

Layered over the bottom of the video, the graphic (Figure 6.1) depicts the activity and polarity of annotations as a stacked graph which is time-aligned to the timeline of the video. Negative annotations are red, positive are green, and neutral are gray. Hovering over the graph shows a detail pane that tracks with the cursor. The detail pane shows the number of contributors to the discussion at that point in the video as well as a stacked bar chart indicating each individual annotation, its polarity, and whether or not it has evidence associated with it. Clicking the graph locks or unlocks the detail pane's position. Clicking on individual annotations on the stacked bar in the detail pane animates and expands a second detail pane with actual tag semantics, comment text for reading, or links to online sources.

## **Experiment**

We were interested in understanding to what extent the visualization we developed could modulate people's credibility evaluations of the video. We conducted an experiment comparing participants' ratings between the experimental video player and a control version which did not have any additional graphics.

Each participant in the study completed a background questionnaire to collect data about political viewpoint, English fluency, and their interest levels in a variety of political issues. The participant was then exposed to three informational videos in succession. A blue and yellow mapping was provided for color blind users. In the experimental condition, participants were told that the video had been evaluated by 8 independent journalists and that these evaluations were accessible via the graphic. Each video was followed by a brief questionnaire designed to assess the reactance to the video overall and to its individual claims in terms of belief. After

the last video, participants in the experimental condition completed a final questionnaire to elicit information about the user experience.

### **Participants and Data Filtering**

Participants were recruited using Amazon Mechanical Turk (AMT), a marketplace where users complete micro tasks for small payments. 104 people participated in the study for \$0.25 each. Because of the low compensation, we found we needed to filter results based on a set of 3 control questions. These control questions, specific to the content of each video rather than general as in [57], were easy to answer correctly if the video was watched. If any of these answers were incorrect, the questionnaire data from that participant was not included in our analysis. Theories of attitude change dictate that reception and comprehension of information are essential [126]. We therefore also filtered data by users' self-reported English fluency (less than 5 on a 7 point scale was excluded). Also, if a participant indicated that they had seen a particular video before, their data was not considered in the analysis of that video since we wanted to focus on the effect of the visualization on people's first exposure to a video. 27% of responses were filtered using the above methods.

As the amount of noise in user ratings can be substantial on AMT [57, 58, 107], in addition to the verifiable content questions we also used interaction logs as a method for observing and filtering data based on participants' degree of engagement with the graphic in the experimental condition. This provided objective information about the level of engagement that participants offered the video and graphic beyond that of answering the content questions correctly.

## **Content Selection and Preparation**

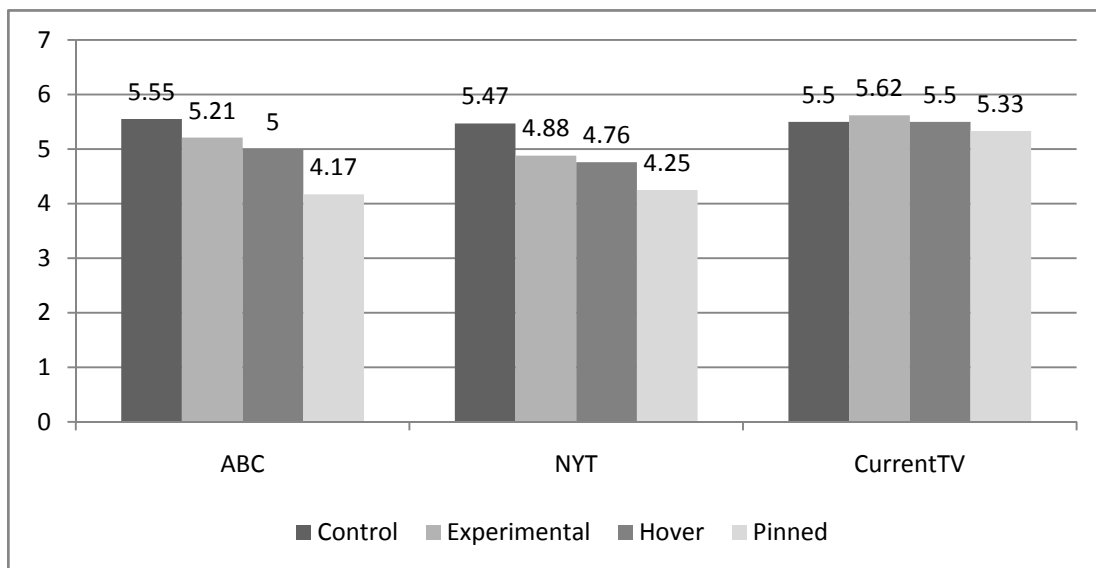
We selected produced news packages because they generally tend to be more visually interesting and information dense than user generated content. The videos were chosen for a diversity of topics and included an ABC Medical Minute podcast about the use of electronic medical records, a NYT video report about traffic and congestion in New York City, and a CurrentTV documentary promo about the Mexican drug war. All videos were cropped and re-encoded to remove any trace of a corporate logo which might confound credibility assessments of the content. The content was annotated by an experimenter on a granular, statement-by-statement basis using a semi-structured information quality ontology. The ABC and NYT videos were annotated mostly negatively with 86% of annotations negative whereas the CurrentTV video was annotated mostly positively with 90% of annotations positive. 24 of the 101 annotations had sources such as web pages and news stories supporting them.

## **Results**

For both of the negatively annotated videos, ABC and NYT, we found there to be an effect of the graphic on credibility ratings, with stronger effects observed for users who engaged the graphic more. To assess the degree of engagement of users we considered whether or not they had pinned the graphic in order to interact with the detail pane (high engagement) or whether or not they had hovered their mouse over the graphic to get more information (medium engagement). If they did neither of these we considered them low engagement users.

The ABC video showed a significant effect for the visualization on the overall credibility of the information in the video ( $F(1, 71) = 11.47, p = .001$ ) for high engagement users. The effect was weak for medium engagement users ( $F(1, 71) = 3.21, p = .07$ ) and there was no effect for low engagement users. For the NYT video the effect was strongest for high engagement users ( $F(1, 72) = 11.11, p = .001$ ), and weaker for medium engagement ( $F(1,72)=5.42, p=.023$ ) or low engagement ( $F(1, 72) = 4.95, p = .029$ ) users. In the case of the CurrentTV video, which was positively annotated, there were no significant effects observed. Means of overall credibility for the different videos and engagement levels are shown in Figure 6.2. Our analysis incorporates covariates collected from the background survey which would be expected to effect belief ratings such as the user’s political viewpoint and topical knowledge and interest [126].

We also considered people’s belief ratings on individual claims within the



**Figure 6.2.** Mean agreement with the statement: “Overall I consider the information presented in the video to be credible.” for each video and across the different conditions and engagement levels.

video. We found significant effects only for high engagement users and only on claims that had been negatively annotated. Claims that had mixed annotations (positive *and* negative) or that were positively annotated saw no effect on their belief ratings. For the ABC video, two of the three negatively annotated claims saw an effect for high engagement users ( $F(1,71) = 4.65, p = .034$  and  $F(1, 71) = 8.09, p = .006$ ). For the NYT video, one of the three negatively annotated claims saw a strong effect ( $F(1, 72) = 9.56, p = .003$ ) and another claim saw a very weak effect ( $F(1, 72) = 3.07, p = .084$ ) for high engagement users.

### User Response

Consistent with what others studying video graphics have found [53] there was a mixed response to having the additional interactive graphics at the bottom of the screen. Several people complained of there being too much information or of being distracted from the video by the graphics. The mean value for agreement with the statement, "I found the information graphics at the bottom of the video distracting" was 3.71 indicating that most people were not largely distracted. However, some people still had distraction problems, and as one person put it, "*It [the graphic] just distracted me and I actually barely looked at it because I was trying to watch the video.*"

Overall most people didn't appear to have too much difficulty with understanding the graphics. The mean for agreement with the statement: "I had difficulty making sense of the information graphics at the bottom of the screen" was 3.83. The mean for agreement with the statement: "I found the information graphics at the bottom of the screen frustrating" was 3.08. These statistics indicate that for

the most part people did not have major problems with comprehending or using the graphics.

But while the graphics weren't overly distracting or incomprehensible, most users would still prefer *not* to have the additional graphics there. The mean agreement rating for the statement: "I liked having the interactive graphic at the bottom of the video." was 3.50 and the mean agreement rating for the statement "I would prefer a video site that had a video player with information graphics similar to the one I saw here." was only 3.33.

On the other hand qualitative feedback does reveal some of the more positive reactions to the graphics. There were several comments that indicated that the graphics and the comments helped in understanding the video and were interesting to see while watching the video. Some thought it provided additional context and appreciated seeing the sources used to back up comments. For instance, one participant wrote, *"It gave the claims in the video more context, and allowed me to interpret what I was hearing better. Claims that I found hard to believe correlated with negative annotations."* Another participant remarked, *"I liked the information it provided, especially the citing and links to sources."* Several users found the granularity of the commenting to be novel. One noted, *"I like that you could see the key areas that comments were based on. That way when someone leaves a comment you can trace it to a specific part of the video and know exactly what they are referencing."*

## Discussion

The strength of the effect of our visualization on credibility evaluations varied with the degree of engagement of users. Users who interacted more with the negative graphics reported more severe credibility ratings. These results are consistent with theories of attitude change and persuasion [35, 126] which predict the importance of information saliency and processing in facilitating attitude change. The more attention users gave to the graphic the more they assimilated what it was saying in terms of the quality of the information in the video.

For the CurrentTV video which was annotated in a majority (90%) positive way there was no observed change in credibility evaluations. Theory predicts that people are more likely to scrutinize disconfirming evidence and to accept confirming evidence at face value [70]. Credibility ratings of the CurrentTV video were already high in the control condition (5.62) therefore annotations which bolstered the prevalent high belief were accepted and not scrutinized. In the ABC and NYT videos, belief ratings in the control condition were also high (5.55 and 5.47 respectively) but the negative annotations were scrutinized to a greater degree because they disagreed with prevalent beliefs.

Our results indicate that saliency and attention to the graphics we created are essential to their impact on users. Saliency could be ratcheted up by using more vivid colors or by incorporating animation to get attention. However, incorporating other methods to make the graphics salient would likely serve to degrade the user experience by increasing distraction. At the same time, presenting annotations that are inconsistent with users' prior beliefs should naturally receive more attention.

This suggests that annotations that contradict prior beliefs will have the most impact when presented in this way.

We operate under the assumption that the creators of the annotations are benign: professionals such as journalists whose ethical standards dictate that they act truthfully and honestly in their assessment of the video. However, there is always a danger in developing persuasive technology since it could be abused or misappropriated. For instance, having an open annotation system could engender gaming and bias of the annotations for personal or institutional gain. In syndication of annotations designers should carefully consider the motivation and self-interests of participants as well as decision factors like *when* to share annotations.

In future work we would also like to assess the role of different types of content. For instance, both Pirolli et al. [92] and Kittur et al. [58] found differences in credibility and trust levels when considering skeptically and controversially categorized content on Wikipedia. More work needs to be done to measure the potential for our method to impact credibility across a broader range of video content.

## CHAPTER 7

### CONCLUSIONS AND FUTURE WORK

This dissertation has explored the design and evaluation of collaborative video annotation and presentation systems as motivated by the desire for better information quality in online media. In this chapter we consider our conclusions and contributions to research that have followed from the design, development, and evaluation of Audio Puzzler, Videolyzer, and Videolyzer CE. We also delineate some of the limitations of our approaches and highlight some areas for future work that have been uncovered in the process of this dissertation.

In Audio Puzzler we showed that game-based methods can be used to turn arbitrary online video into puzzles that allow for the collection of highly accurate time-stamped transcript metadata for video. While other games in this vein of human computation have been used for image and music tagging, they oftentimes rest on the same mechanic: consensus between pairs. The Audio Puzzler game design contributes an existence proof that a puzzle based game mechanic, played solo, can also be effective. Internal consistency is enforced by the puzzle pieces themselves rather than through external consensus. We then showed that by having users implicitly collaborate, by aggregating the results of their puzzle completions, we could enhance the data quality of the collected transcripts to 99% accuracy. We contribute algorithms to aggregate and improve the overall accuracy of the transcripts. Finally, we contribute a deeper understanding of the effects of content in creating games out of previously authored content and an analysis of contextual effects and player knowledge on the efficacy of human computation.

In the process of developing Videolyzer we distilled the essential facets of information quality which we thought should be embodied in a simplified interface incorporating a set of quality tags and related interactions. Through our laboratory evaluation of Videolyzer we found this semi-structured approach toward annotation to be usable and viable as it did not significantly detract from users' ability to communicate and express themselves. The information quality ontology and interactions that we developed constitute a design contribution that could be implemented by others interested in similar applications.

We were interested in studying the usefulness of several video annotation features implemented in Videolyzer including: transcript integration, granularity, suggestive automation, and integration of an argumentation system based on our model. Our evaluation of Videolyzer contributes knowledge about the efficacy of this range of video annotation interface features. Our results show that the transcript integration and coupling to the video interactions was usable and preferable as a method for navigating and annotating video. Designers of video analytic interfaces should strongly consider the familiarity of text as an interaction index for video when the content makes this possible. Users also liked the granularity that the interface afforded as they felt it helped them focus the discussion. While adding some additional interaction overhead to the interface, granularity is one mechanism that functions to facilitate the scaling up of large online discussion systems. We found a low to moderate user benefit associated with the automatically extracted markers we provided. Designers should carefully consider the benefits of integrating such automation since we didn't find that the value provided was

necessarily worth the additional complexity added to the interface. Finally, we found that in the process of using Videolyzer to do video analysis users' understanding of comprehensiveness, multiple perspectives, context, and the quality of the video were enhanced.

While Videolyzer was explicitly designed for engaged and motivated users such as bloggers and journalists we felt that the value of the annotations collected should be shared with a wider audience. We contribute the design of a video player which syndicates the annotations collected with Videolyzer into a simplified presentation of information quality cues shown in line with the video. We evaluated this interface, Videolyzer CE, and found that graphics showing annotation activity, polarity, and sources could influence people's perceptions of the credibility of the video. We found there to be some limitations associated with the approach, such as limited impact on unengaged viewers and strong content effects which dictate the efficacy of annotations on changing people's credibility evaluations. While consistent with theoretical predications in attitude change, we contribute a prototype and evaluation showing the real potential of such interfaces to syndicate annotations and manipulate the perception of online video.

### **Limitations**

Here we consider some of the more significant limitations associated with the annotation approaches we have developed. We also reflect on the implications and generalizability of our results.

## **Temporal Information Dynamics**

From the standpoint of information quality, one of the limitations of a collaborative annotation approach is that it can only be effective *after* something has been published. This means, for instance, that bad information can diffuse until it has been marked as poor quality. The approach forms a post-filter, but lacks a pre-filter - the ability to screen poor quality information before it is published. In general, if the information quality filtering activity (annotation) is not centrally coordinated, annotations may also be slow to accrete or irregularly and sparsely distributed. This could lead to significant delays in aggregating annotations for a particular piece of media depending on the number of annotators participating in the collaborative endeavor. In the case of the video transcripts for Videolyzer there will be some delay in the aggregation of the 5 to 8 transcripts from Audio Puzzler depending on the amount of traffic there. One way to reduce this delay time is to post audio puzzles to Amazon Mechanical Turk, though this costs money.

## **Content Suitability**

Across both annotation systems studied we found there to be interesting content considerations. In the case of Audio Puzzler engaging content such as from comedies is more fun to play the game with. Information content such as news or documentary may appeal to a smaller audience and boring or slow audio such as meeting recordings is unlikely to be a compelling source of content for the game. Even though an arbitrary piece of audio could theoretically be used to create a puzzle for the game, in practice, some editorial decision needs to be made by a human to, at the very least, select appealing content.

Another aspect of content to consider in the context of Audio Puzzler is the need for privacy or confidentiality of the video being used in the game. For sure there is plenty of video that is publicly available and could be turned into puzzles that are available on the internet, but there is also a substantial amount of proprietary video that could also be enhanced through transcription. The problem is that making this private video available through the game might not be desirable or even legal in some situations.

There are also several content considerations worth noting for Videolyzer. As a collaborative tool Videolyzer is of maximal benefit for video content that can be interpreted in many different ways (i.e. it is polysemous and open to interpretation from different perspectives). For instance, complex issues which involve many different perspectives or stakeholders are likely to benefit from the re-contextualization facilitated through multiple collaborators contributing to Videolyzer. On the other hand, our results in Videolyzer CE indicate that the effects of syndicating those annotations will only be felt in situations where there is consensus about the quality of a particular claim. Claims with both positive and negative annotations saw no effect on their credibility rating. Sections of video that are unanimously annotated in a particular direction will see the most benefit in Videolyzer CE. This combination of effects suggests that for maximal influence, only *unanimous* annotations should be syndicated in Videolyzer CE. This may make sense since one would only want to perpetuate an interpretation of the video if there were sufficient consensus to warrant that. On the other hand, the highly

motivated users of Videolyzer also found significant value to seeing the complexity, comprehensiveness, and perspectives available there.

In many ways the issue of content suitability described here echoes the conception of low- and high-feedback information discussed by Gentzkow and Shapiro [40]. Low-feedback information involves those things that are complex or distant from direct experience and whose outcomes are difficult to observe. On the other hand high-feedback information includes things that are concrete and whose outcomes are immediately observable, such as a weather forecast. In this terminology, the re-contextualization afforded by Videolyzer is most suited for low-feedback information; users can contribute and be exposed to the variety of perspectives on an issue. At the same time, annotations on high-feedback information may be most suited to syndication through Videolyzer CE since it is unlikely that unanimity of evaluation would be achieved for low-feedback information.

### **Media Literacy and Knowledge of Users**

In any human computation system whether that be a game based one as in Audio Puzzler, or a formal annotation system as in Videolyzer, the media literacy and knowledge of the individual users plays a role in the value of the information product. In Audio Puzzler we found a slight correlation between the number of years spent in an English speaking country and the accuracy of transcripts produced by a user. Some of this effect may in fact be attributed to language proficiency, but another aspect is cultural knowledge. In some clips there were references to pop culture such as the names of bands, singers, or products, that may be more difficult

for players to recognize and transcribe if they did not have cultural familiarity with those named entities.

For Videolyzer, media literacy and the ability to look for and recognize different facets of information quality is essential. By controlling our evaluation of Videolyzer and recruiting only people who were known bloggers or journalists we sidestepped this issue for the most part. In an online context where expertise and credentials of individuals cannot be verified a priori other methods from reputation systems would have to be employed to build models of user behavior over time. However, opening the system up also exposes it to less informed users injecting confusion into the discussion or at worst gaming it to intentionally produce a bias.

In general, the users of a system like Videolyzer, whether it is presented online for all or is used only internally within a single newsroom will have a variety of backgrounds and knowledge. This distribution of the quality management process across many individuals fits nicely with Walter Lippmann's observation that people can contribute in areas where they have particular expertise, experience, or knowledge and reap the benefits of others' contributions if they themselves are not knowledgeable [69].

## **Broader Implications and Impact**

### **Annotation Systems**

Games have begun to sprout up in all kinds of different domains to motivate the participation of people in solving tasks that cannot yet be automated. Audio Puzzler is a step in this direction and shows the power of aggregating the

contributions of many people together using some known structure. The future of annotation systems probably more squarely fits at the intersection of automatic and manual efforts however. Just as aggregation of human effort can increase data quality for metadata like transcripts, it can likewise be aggregated and merged together with computer effort. The challenge is in identifying systems that can incorporate some amount of computer output together with human processing capabilities (such as that motivated through games) in order to achieve high metadata accuracy. This will lead to more scalable solutions that are not as labor intensive. There remains much work to be done in game studies in order to identify and harness different formal game mechanics which can incorporate human computation and produce a viable and useful output. As these mechanics are uncovered, studied, and applied to different applications together with automation, annotation quality will be enhanced and new applications facilitated.

As was shown in the synergy between Audio Puzzler and Videolyzer, quality metadata enables new applications and interaction paradigms. In our case the time-stamped transcripts from Audio Puzzler resulted in an interaction experience for video annotators in Videolyzer that was familiar and intuitive based on their experience in manipulating text. Furthermore, the granular application of a semi-formal argumentation system has implications not only for video annotation, but for any form of anchored media. The argumentation system allows people to express themselves and expand the context of the underlying media; a value that can apply across image, video, text, or interactive information graphics. In the same way as transcript metadata is used to enhance Videolyzer, the information quality metadata

produced by Videolyzer can be used to enable persuasive applications such as Videolyzer CE. In this way, we envision chains of semantic data flowing between applications. The types of applications that will be enabled by large efforts of information tagging such as the Semantic Web will have a profound impact on the way that we author and consume information and media in the future.

### **Influence and Persuasion**

The communicational and ethical implications of Videolyzer and Videolyzer CE are profound if we consider the potential for the technology to either correctly inform people or, if abused, to mislead people further. At the outset of this dissertation the premise was that through aggregated information quality evaluations people would become better informed about the video information they were consuming online. This premise holds for both Videolyzer and Videolyzer CE if we also assume that the goals and motivations of those providing the annotations are pure. Given this assumption, professional watchdogs operating under ethical guidelines that would be familiar to journalists would be able to spread a better quality of information, and have their evaluations modulate credibility perceptions of some of the worst quality information found online.

On the other hand, if misappropriated, Videolyzer CE could serve to syndicate the annotations of those interested primarily in manipulating others for personal, ideological, or financial gain. Essentially, the annotations added to something like Videolyzer are no different than anything else published by someone: they can be biased, misinformed, inaccurate, or unreliable and only over time should the system be expected to identify this poor information. While users of

Videolyzer found value in seeing the varied perspectives of complex issues, some thought should be put into *when* these annotations should be syndicated into something like Videolyzer CE. Should this only happen when a certain consensus threshold is met? What if new information is uncovered which drastically changes the evaluation of a particular claim? These are questions that would need answering if such a system were to be implemented in a real news ecosystem.

### **Information Access**

One of the promises of rich annotation mechanisms is the enhanced searchability of media afforded by them. Both Videolyzer and Audio Puzzler anchor annotations to time segments of video which allows for searching within the time stream of annotations for keywords. The implications of these approaches for media search and information retrieval are that people will spend less time looking for the section of video that they are most interested in watching. At the same time, models of user behavior could be combined with annotations in the system to provide personalized recommendations for annotations that are of most interest to an individual.

### **Research Opportunities**

Games which incorporate re-mediations of existing content are an exciting area of research that has only really been scratched at by Audio Puzzler. In particular, automatic media analysis mechanisms such as computer vision and computational linguistics can be used to transform existing authored media such as news stories, audio podcasts, or video broadcasts into games. This type of game

could be used to expose players to content that they otherwise might not be exposed to, thereby providing raised awareness for news information concomitant to play. This research could also incorporate controlled user studies to examine the communicational impact of such games. The implications for the communication of important information such as health and news to game savvy audiences could be substantial.

The general principles of information quality described in this dissertation could also be applied and extended to other media such as audio, text, or even information graphics. Information graphics in particular have challenges associated with the fact that they can represent models or data that may be incomplete, uncertain, or procedural. Can a similar application of annotation be used to enhance the credibility of information graphics? We would like to look into methods for annotating info graphics and then presenting those annotations as a way for providing transparency and accountability to the graphics.

As elucidated above, the knowledge of users participating in human computation systems is oftentimes a defining factor in the value of the output produced. This begs the question of better understanding the background, context, and culture of those participating in these systems. As platforms like Amazon Mechanical Turk gain traction as markets for accomplishing human computation it will be important to understand the motivations, demographics, cultures, and processing limitations of participants there. We plan to do a detailed study of these factors specifically for AMT since this will help inform its use in a range of applications.

In this dissertation we have studied a range of video annotation methods, but there are still open questions. In particular we found our results concerning the integration of automatic suggestions mixed, however this could be due to a constrained context or to short videos. Future research should look at the role of suggestive interfaces for longer videos and under different circumstances of use. As always, careful thought needs to be put into finding the intersection of things that can be reliably suggested automatically and the things that people actually care to be made aware of via suggestions.

We also believe there is a huge potential to develop smart discussion and commenting systems which can build off of some of the ideas presented here. In particular, as commenting systems grow to include tens of thousands of comments (as is already happening on some stories at the New York Times website), better methods of organizing those comments will be needed to ease consumption and sensemaking. In this work we developed a structured set of tags that can be used to organize the discussion, however, other methods based more on textual analysis and machine learning could also be employed to provide structured interfaces. Granularity of commenting was perceived by some to provide additional focus to the discussion but a deep study of this effect remains to be run. If proven effective, such a strategy could easily be incorporated by other online commenting systems to reduce the overall burden of consuming comments that are of most interest and relevance to a user.

## REFERENCES

1. Compendium, 2008.
2. Merriam-Webster Online.
3. Open Calais., Accessed 9/8/2008.
4. Abowd, G., Gauger, M. and Lachenmann, A., The Family Video Archive: an annotation and browsing environment for home movies. in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, (2003), 1-8.
5. Ames, M. and Naaman, M., Why we tag: motivations for annotation in mobile and online media. in *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, (2007), ACM, 971-980.
6. Anderson, C. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.
7. boyd, d., Lee, H.-Y., Ramage, D. and Donath, J., Developing Legible Visualizations for Online Social Spaces. in *HICSS*, (2002).
8. Brill, E., A Simple Rule-Based Part of Speech Tagger. in *Conference on Applied Computational Linguistics*, (1993), 112-116.
9. Buckingham Shum, S., Sensemaking on the Pragmatic Web: A Hypermedia Discourse Perspective. in *Proceedings of the International Conference on the Pragmatic Web*, (2006).
10. Burr, B., VACA: A Tool for Qualitative Video Analysis. in *Extended Abstracts of CHI*, (Montreal, Canada, 2006).
11. Casares, J., Long, A.C., Myers, B., Bhatnagar, R., Stevens, S., Dabbish, L., Yocum, D. and Corbett, A., Simplifying video editing using metadata. in *Proceedings of DIS*, (2002).
12. Chklovski, T., Ratnakar, V. and Gil, Y., User Interfaces with Semi-Formal Representation: a Study of Designing Argumentation Structures. in *Intelligent User Interfaces (IUI)*, (2005).
13. Civan, A. and Pratt, W., Supporting Consumers by Characterizing the Quality of Online Health Information: A Multidimensional Framework. in *Proceedings of HICSS*, (2006).
14. Cockburn, A. and Dale, T., CEVA: A Tool for Collaborative Video Analysis. in *GROUP*, (1997), 47-55.

15. Correia, N. and Chambel, T., Active Video Watching using Annotation. in *ACM Multimedia*, (Orlando, FL, 1999), 151-154.
16. Costa, M., Correia, N. and Guimarães, N., Annotations as Multiple Perspectives of Video Content. in *Proceedings of ACM Multimedia*, (2002), 283-286.
17. Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*. Harper Perennial, 1991.
18. Cui, J., Wen, F., Xiao, R., Tian, Y. and Tang, X. EasyAlbum: an interactive photo annotation system based on face clustering and re-ranking *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, San Jose, California, USA, 2007.
19. Davis, M. Media Streams: An Iconic Visual Language for Video Representation. in Baecker, R.M., Grudin, J., Buxton, W. and Greenberg, S. eds. *Readings in Human-Computer Interaction: Toward the Year 2000*, Morgan Kaufmann, 1995, 854-866.
20. Deci, E. and Flaste, R. *Why We Do What We Do: Understanding Self-Motivation*. Penguin, 1995.
21. Desurvire, H., Caplan, M. and Toth, J., Using heuristics to evaluate the playability of games. in *Proceedings of CHI*, (2004), 1509-1512.
22. Diakopoulos, N. and Chiu, P., PhotoPlay: A Collocated Collaborative Photo Tagging Game on a Horizontal Display. in *Proceedings addendum of UIST*, (2007).
23. Diakopoulos, N. and Essa, I., An Annotation Model for Making Sense of Information Quality in Online Video. in *International Conference on the Pragmatic Web*, (Uppsala, Sweden, 2008).
24. Diakopoulos, N. and Essa, I., Mediating photo collage authoring. in *Proceedings of the 18th annual ACM symposium on User interface software and technology (UIST)*, (2005), 183-186.
25. Diakopoulos, N. and Essa, I., Videotater: an approach for pen-based digital video segmentation and tagging. in *Proceedings of UIST*, (2006), ACM, 221-224.
26. Diakopoulos, N., Essa, I. and Jain, R., Content Based Image Synthesis. in *Proceedings of Conference on Image and Video Retrieval (CIVR)*, (2004).
27. Diakopoulos, N., Goldenberg, S. and Essa, I., Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists. in *Conference on Human Factors in Computing Systems (CHI)*, (Boston, MA, 2009).

28. Diakopoulos, N., Luther, K. and Essa, I., Audio Puzzler: Piecing Together Time-Stamped Speech Transcripts with a Puzzle Game. in *ACM Multimedia*, (2008).
29. Dieberger, A., Dourish, P., Höök, K., Resnick, P. and Wexelblat, A. Social Navigation: Techniques for Building More Usable Systems *interactions*, 2000.
30. Dix, A., Finlay, J., Abowd, G. and Beale, R. *Human-Computer Interaction*. Prentice Hall, 2003.
31. Drucker, S., Wong, C., Roseway, A., Glenner, S. and De Mar, S., MediaBrowser: reclaiming the shoebox. in *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, (2004), ACM, 433-436.
32. Evans, J. *Bias in Human Reasoning: Causes and Consequences*. Lawrence, 1989.
33. Few, S. *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media, 2006.
34. Fogg, B.J. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan, 2003.
35. Fogg, B.J., Prominence-interpretation theory: explaining how people assess credibility online. in *CHI '03: CHI '03 extended abstracts on Human factors in computing systems*, (2003), ACM, 722-723.
36. Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. and Treinen, M., What makes Web sites credible?: a report on a large quantitative study. in *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, (2001), ACM, 61-68.
37. Fogg, B.J. and Tseng, H., The elements of computer credibility. in *Proceedings of CHI*, (1999), 80-87.
38. Franklin, B., Hamer, M., Hanna, M., Kinsey, M. and Richardson, J.E. *Key Concepts in Journalism Studies*. Sage Publications, 2005.
39. Franzen, R. Warning Letter, 2008.
40. Gentzkow, M. and Shapiro, J.M. Media Bias and Reputation. *Journal of Political Economy*, 114 (2). 280-316.
41. Gil, Y. and Ratnakar, V., Trusting Information Sources One Citizen at a Time. in *International Semantic Web Conference (ISWC)*, (2002).
42. Giles, J. Internet Encyclopaedias go head to head *Nature*, 2005.

43. Gillmor, D. *We the Media: Grassroots Journalism by the people, for the people*. O'Reilly, 2006.
44. Goldstein, T. *Journalism and Truth: Strange Bedfellows*. Northwestern University Press, 2007.
45. Hagedorn, J., Hailpern, J. and Karahalios, K.G., VCode and VData: Illustrating a New Framework for Supporting the Video Annotation Workflow. in *Advanced Visual Interfaces (AVI)*, (Naples, Italy, 2008), 317-321.
46. Hassenzahl, M., Beu, A. and Burmester, M. Engineering Joy. *IEEE Software* (Jan / Feb). 70-76.
47. Haubold, A., Dutta, P. and Kender, J.R., Evaluation of Video Browser Features and User Interaction with VAST MM. in *ACM Multimedia*, (Vancouver, Canada, 2008), 449-458.
48. Haubold, A. and Kender, J.R., Alignment of Speech to Highly Imperfect Text Transcriptions. in *International Conference on Multimedia and Expo (ICME)*, (2007), 224-227.
49. Heer, J., Viegas, F. and Wattenberg, M., Voyagers and voyeurs: supporting asynchronous collaborative information visualization. in *Proceedings of CHI*, (2007), 1029-1038.
50. Hoffmann, M.H.G., Logical Argument Mapping: A cognitive-change-based method for building common ground. in *Conference on the Pragmatic Web*, (2007), 41-47.
51. Huang, X., Acero, A. and Hon, H.W. *Spoken Language Processing*. Prentice Hall, 2001.
52. Jackson, B. and Jamieson, K.H. *unSpun: finding facts in a world of disinformation*. Random House Trade Paperbacks, 2007.
53. Josephson, S. and Holmes, M.E., Clutter of Content? How on-screen enhancements affect how TV viewers scan and what they learn. in *Proc. Symposium on Eye Tracking Research & Application*, (2006), 155-162.
54. Keelan, J., Pavri-Garcia, V., Tomlinson, G. and Wilson, K. YouTube as a source of Information on Immunization: A Content Analysis. *Journal of the American Medical Association*, 298 (21). 2482-2284.
55. Kim, P., Eng, T., Deering, M.J. and Maxfield, A. Published criteria for evaluating health related web sites: review. *BMJ*, 318 (March). 647-649.

56. Kirschner, P., Buckingham Shum, S. and Carr, C. (eds.). *Visualization Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer-Verlag, 2003.
57. Kittur, A., Chi, E.H. and Suh, B., Crowdsourcing User Studies With Mechanical Turk in *Proceedings of CHI*, (2008), 453-456.
58. Kittur, A., Suh, B. and Chi, E.H. Can You Ever Trust a Wiki? Impacting Perceived Trustworthiness in Wikipedia *CSCW*, San Diego, CA, 2008, 477-480.
59. Koster, R. *A Theory of fun for game design*. Paraglyph, 2005.
60. Kovach, B. and Rosenstiel, T. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. Three Rivers Press, 2007.
61. Kovach, B. and Rosenstiel, T. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. Three Rivers Press, 2001.
62. Kuhn, T. *The Structure of Scientific Revolutions*, 1996.
63. Kustanowitz, J. and Shneiderman, B. Motivating Annotation for Personal Digital Photo Libraries: Lowering Barriers while Raising Incentives, University of Maryland, 2005.
64. Lampe, C. and Garrett, R.K., It's all News to Me: The Effect of Instruments on Ratings Provision. in *Proceedings of HICSS*, (2007).
65. Lawrence, G. and Grey, D. Subjective Inaccuracies in Local News Reporting. *Journalism Quarterly*, 46. 753-757.
66. Leech, G. *Meaning and The English Verb*. Pearson Education Limited, 2004.
67. Lerman, K. Social Information Processing in News Aggregation. *IEEE Internet Computing*. 16-28.
68. Li, F., Gupta, A., Sanocki, e., He, L.-w. and Rui, Y., Browsing Digital Video. in *Proceedings of CHI*, (2000).
69. Lippmann, W. *Public Opinion*. Free Press, 1922.
70. Lord, C.G., Ross, L. and Lepper, M.R. Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology*, 37 (11). 2098-2109.
71. Luo, H., Fan, J., Yang, J., Ribarsky, W. and Satoh, S.i., Exploring Large-Scale Video News via Interactive Visualization. in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, (2006).

72. Macgregor, G. and McCulloch, E. Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review*, 55 (5).
73. Machill, M., Beiler, M. and Schmutz, J. The Influence of Video News Releases on the Topics Reported in Science Journalism. *Journalism Studies*, 7 (6). 869-888.
74. Mackay, W.E. EVA: an experimental video annotator for symbolic analysis of video data. *SIGCHI Bull.*, 21 (2). 68-71.
75. Maier, S.R. Accuracy Matters: A Cross-Market Assessment of Newspaper Error and Credibility. *Journalism and Mass Communication Quarterly*, 82 (3). 533-551.
76. Malone, T., Heuristics for designing enjoyable user interfaces: Lessons from computer games. in *Proceedings of CHI*, (1982), ACM, 63-68.
77. Manjoo, F. *True Enough: Learning to Live in a Post-Fact Society*. John Wiley & Sons, New Jersey, 2008.
78. Marlow, C., Naaman, M., Boyd, D. and Davis, M., HT06, tagging paper, taxonomy, Flickr, academic article, to read. in *Proceedings of Hypertext and Hypermedia*, (2006), 31-40.
79. McCroskey, J. The effects of evidence in persuasive communication. *The Journal of Western Speech Communications*, 31. 189-199.
80. Meyer, P. Defining and Measuring Credibility of Newspapers: Developing an Index. *Journalism Quarterly*, 65. 567-574.
81. Miller, A. and Edwards, W.K., Give and take: a study of consumer photo-sharing culture and practice. in *Proceedings of CHI*, (San Jose, CA, 2007), 347-356.
82. Miller, G. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63. 81-97.
83. Mills, M., Cohen, J. and Wong, Y., A magnifier tool for video data. in *Proceedings of CHI*, (1992), ACM, 93-98.
84. Mobbs, D., Greicius, M., Abdel-Azim, E., Menon, V. and Reiss, A. Humor Modulates the Mesolimbic Reward Centers. *Neuron*, 40. 1041-1048.
85. Morville, P. and Rosenfeld, L. *Information Architecture for the World Wide Web*. O'Reilly, 2006.

86. Munteanu, C., Baecker, R. and Penn, G., Collaborative Editing for Improved Usefulness and Usability of Transcript-Enhanced Webcasts. in *CHI*, (Florence, Italy, 2008), 373-382.
87. Munteanu, C., Baecker, R., Penn, G., Toms, E. and James, D., The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives. in *Proceedings of CHI*, (2006), 493-502.
88. Nakamura, S., Shimizu, M. and Tanaka, K., Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips? in *Workshop on Information Credibility on the Web (WICOW)*, (2008), 59-62.
89. Nov, O., Naaman, M. and Ye, C., What Drives Content Tagging: The Case of Photos on Flickr. in *CHI '08: Proceedings of the SIGCHI conference on Human factors in computing systems*, (Florence, Italy, 2008), 1097-1100.
90. Pinelle, D., Wong, N. and Stach, T., Heuristic Evaluation for Games: Usability Principles for Video Game Design. in *Proceedings of CHI*, (2008), 1453-1462.
91. Pioch, N. and Everett, J., POLESTAR: Collaborative Knowledge Management and Sensemaking Tools for Intelligence Analysts. in *Conference Information and Knowledge Management*, (2006), 513-521.
92. Pirolli, P., Wollny, E. and Suh, B., So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility. in *Proc. CHI*, (2009), 1505-1508.
93. Ramos, G. and Balakrishnan, R., Fluid interaction techniques for the control and annotation of digital video. in *Proceedings of UIST*, (2003), 105-114.
94. Rieh, S. Judgment of Information Quality and Cognitive Authority in the Web. *Journal of the American Society for Information Science and Technology*, 53 (2). 145-161.
95. Rupar, V. How Did You Find That Out? Transparency of the Newsgathering Process and the Meaning of News. *Journalism Studies*, 7 (1). 127-143.
96. Ryall, K., Li, Q. and Esenther, A., Temporal Magic Lens: Combined Spatial and Temporal Query and Presentation. in *Proceedings of INTERACT*, (2005).
97. Schiff, S. Know It All: Can Wikipedia conquer expertise? *The New Yorker*, 2006.
98. Sen, S., Lam, S., Rashid, A., Cosley, D., Frankowski, D., Osterhouse, J., Harper, M. and Riedl, J., tagging, communities, vocabulary, evolution. in *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, (2006), ACM, 181-190.

99. Serano, B., Buckingham Shum, S. and Motta, E., ClaimSpotter: An Environment to Support Sensemaking with Knowledge Triples. in *Intelligent User Interfaces (IUI)*, (2005).
100. Shipman, F. and Marshall, C., Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. in *CSCW*, (1999), 333-352.
101. Shneiderman, B. Designing for fun: how can we design user interfaces to be more fun? *interactions*, 11 (5). 48-50.
102. Shneiderman, B. and Kang, H., Direct Annotation: a drag-and-drop strategy for labeling photos. in *Proceedings of the International Conference on Information Visualization*, (2000).
103. Simon Buckingham Shum, R.S., Michael Daw, Ben Juby, Andrew Rowley, Michelle Bachler, Clara Mancini, Danius Michaelides, Rob Procter, David De Roure, Tim Chown, Terry Hewitt. Memetic: An Infrastructure for Meeting Memory, 2006.
104. Singer, N. A Birth Control Pill That Promised Too Much *New York Times*, 2009.
105. Smeulders, A., Worring, M., Santini, S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (12). 31.
106. Smith, M.A. and Fiore, A.T., Visualization Components for Persistent Conversations. in *Proc CHI*, (2001), 136-143.
107. Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks *Conference on Empirical Methods in Natural Language Processing*, 2008.
108. St.Clair, A., Fong, M., Diakopoulos, N. and Essa, I., NARC: The News Article Revision Comparator. in *Proceedings addendum of User Interface Software Technology (UIST)* (Newport, Rhode Island, 2007).
109. Stark, L., Whittaker, S. and Hirschberg, J., ASR Satisficing: The effects of ASR accuracy on speech retrieval. in *Proceedings of International Conference on Spoken Language Processing*, (2000).
110. Strong, D., Lee, Y. and Wang, R. Data Quality in Context. *Communications of the ACM*, 40 (5). 103-110.

111. Stvilia, B., Twidale, M., Smith, L. and Gasser, L. Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59 (6). 983-1001.
112. Suh, B., Chi, E.H., Kittur, A. and Pendleton, B.A., Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashboard. in *Proc. CHI*, (2008), 1037-1040.
113. Sweetser, P. and Wyeth, P. GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3 (3). 3-3.
114. Taylor, R. *Value-Added Processes in Information Systems*. Ablex, 1986.
115. Turnbull, D., Liu, R., Barrington, L. and Lanckriet, G., A Game-Based Approach for Collecting Semantic Annotations of Music. in *International Symposium on Music Information Retrieval (ISMIR)*, (2007).
116. Vasconcelos, N. From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval. *IEEE Computer*, 40 (7). 20-26.
117. Viégas, F.B., Wattenberg, M., Ham, F.v., Kriss, J. and McKeon, M., Many Eyes: A Site for Visualization at Internet Scale. in *IEEE InfoVis*, (2007).
118. Volkmer, T., Smith, J. and Natsev, A., A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. in *Proceedings of ACM Multimedia*, (2005), 892-901.
119. von Ahn, L. and Dabbish, L., Labeling images with a computer game. in *Proceedings of CHI*, (2004), 319-326.
120. von Ahn, L., Ginosar, S., Kedia, M., Liu, R. and Blum, M., Improving accessibility of the web with a computer game. in *Proceedings of CHI*, (2006), ACM, 79-82.
121. von Ahn, L., Kedia, M. and Blum, M., Verbosity: a game for collecting common-sense facts. in *Proceedings of CHI*, (Montréal, Canada, 2006), 75-78.
122. von Ahn, L., Liu, R. and Blum, M., Peekaboom: A game for locating objects in images. in *Proceedings of CHI*, (Montréal, Canada, 2006), 55-64.
123. von Ahn, L., Maurer, B., McMillen, C., Abraham, D. and Blum, M. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321. 1465-1468.
124. Wactlar, H., Hauptmann, A., Christel, M., Houghton, R. and Olligschlaeger, A. Complementary video and audio analysis for broadcast news archives. *Commun. ACM*, 43 (2). 42-47.

125. Weber, K. and Poon, A., Marquee: a tool for real-time video logging. in *Proceedings of CHI*, (1994), ACM, 58-64.
126. Wyer, R.S., Jr., and Albarracín, D. Belief formation, organization, and change: Cognitive and motivational influences. in D. Albarracín, B. T. Johnson and Zanna, M.P. eds. *The handbook of attitudes*, Erlbaum, 2005.
127. Yamamoto, Y., Nakakoji, J. and Akio, T., The Landscape of Time-Based Visual Presentation Primitives for Richer Video Experience. in *INTERACT*, (2005).
128. Yang, Y., Zhu, B.B., Guo, R., Yang, L., Li, S. and Yu, N., A Comprehensive Human Computation Framework - With Application to Image Labeling. in *ACM Multimedia*, (Vancouver, Canada, 2008), 479-488.
129. Yetim, F., DISCOURSIUM for Cooperative Examination of Information in the Context of the Pragmatic Web. in *Conference on the Pragmatic Web*, (2007).
130. Yin, R. *Case Study Research : Design and Methods (Applied Social Research Methods)*. SAGE Publications, 2002.
131. Yuan, J., Liberman, M. and Cieri, C., Towards an Integrated Understanding of Speaking Rate in Conversation. in *Proceedings of the Conference on Spoken Language Processing*, (2006).

## **VITA**

### **Nicholas A. Diakopoulos**

Nicholas Diakopoulos hails from the vast state of Rhode Island where he was born and spent the first twenty-two years of his life. After growing up in the colonial town of East Greenwich he shipped off to the grand state capital, Providence, to study at Brown University. While at Brown he studied Computer Engineering and Multimedia Signal Processing, in which he received his Sc.B. in 2002. After graduating from Brown he moved to Europe, first to England and then to Germany to work in Darmstadt at the Fraunhofer Institut für Graphische Datenverarbeitung. He returned to the US in 2003 and began his graduate studies in the College of Computing at the Georgia Institute of Technology where he received his MS in Computer Science in 2005. While working on his PhD he had the opportunity to do summer internships at Joanneum Research IIS and Fuji-Xerox Palo Alto Laboratory and played a major role in the development of the course and symposium on Computational Journalism at Georgia Tech.