Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination

Dhevi J. Rajendran
Computer Science, Rice University

Andrew T. Duchowski
School of Computing, Clemson University

Pilar Orero
Ctr. for Rsrch. in Ambient Intelligence and Accessibility of Catalonia
Universitat Autònoma de Barcelona

Juan Martínez
TransMedia Catalonia

Pablo Romero-Fresco
Media, Culture, Language, Roehampton University

May 1, 2011

Abstract

Our work serves as an assay of the visual impact of text chunking on live (respoken) subtitles. We evaluate subtitles constructed with different chunking methods to determine whether segmentation influences comprehension or otherwise affects the viewing experience. Disparities in hearing participants' recorded eye movements over four styles of subtitling suggest that chunking reduces the amount of time spent reading subtitles.

1 Introduction

With the rapid proliferation of digital media content (particularly video, e.g., on mobile phones, YouTube™, etc.), subtitles have become a popular means of augmenting audio/video streams for numerous purposes. Subtitles are a means of content accessibility for the Deaf or Hard-Of-Hearing, an aid to language acquisition,[1] and a source of entertainment for those spoofing the original video content. An online subtitling subculture has emerged, involved in the creation of subtitle content as well as in the development and maintenance of subtitling software. Several popular subtitle formats

[1] According to the U.S. Federal Communications Commission, or FCC, intralingual language captions improve comprehension and fluency.

are available for constructing subtitle streams, including SubRip (`.srt`) and Structured Subtitle Format (`.ssf`). Sophisticated solutions have been developed to extract embedded subtitles from feature films and a large number of feature film subtitle “scores” are freely available on the web. Most of these forms of subtitle content creation are offline, the end product usually packaged in a video file format (e.g., MPEG-4) containing video, audio, and subtitle streams.

In contrast, during live subtitle transmission (e.g., generally referred to as *closed captions* in the U.S.), stenographers transcribe stenographic input for display as captions within 2-3 seconds of the representing audio. Recent developments, largely in Europe, involve subtitlers using speech recognition software to generate the captions by listening to the programme and *respeaking* subtitle content (usually a condensed representation). Respeaking can be considered a variation of shadowing—a paced, auditory tracking task involving immediate vocalization of auditory stimuli in the same language, using word-for-word repetition (Lambert, 1989; Boulianne et al., 2010). Respeaking, however, does not exclusively rely on word-for-word repetition (Ribas & Romero-Fresco, 2008). Rather, it relies on the recoding and reproduction of auditory input into successive idea units (Chafe, 1980). Following principles of language processing, in particular, *text chunking*, it has been proposed that subtitles should be visually segmented using punctuations or phrases to delineate visual segments (see Fig. 1) (Martínez & Linder, 2010).

[Figure 1 about here.]

Text chunking, or grouping of a block of text into coherent segments, has been shown in other contexts, such as studies of short-term memory, to increase short-term information retention and speed of information processing (Miller, 1956; Baddeley, 2003). In Human-Computer Interaction, chunking has been used to argue for the design of interface pragmatics to accelerate the acquisition of expert operational skills, e.g., chunking interaction dialogue into meaningful units (Buxton, 1986). Reducing the number of chunks is thought to facilitate the organization and recall of information (Badre, 1982). As an example, an eye tracking study was used to motivate the creation of contact points to represent a co-reference to segmented animation promoting chunking for less effort (Faraday & Sutcliffe, 1999). Although in this instance video was segmented for ease of comprehension, to our knowledge the effect of segmented subtitles has not yet been explored.

European and American subtitling standards establish recommendations for font type, size, color, position, display speed (reviewed below), however, little emphasis has been placed on the subtitles’ linguistic composition. Although extensive research has already been performed regarding subtitle positioning, font size and color, display speed, and their appearance on the screen (Matamala & Orero, 2010), most of this work has, to a large extent, been qualitative.

In this paper we examine the effect of the visual segmentation of subtitles on patterns of eye movements of hearing participants. We present results of an eye-tracking study testing text chunking’s effectiveness when applied to a simulation of respoken subtitles. Four text subtitle styles are evaluated for their impact on both speed at which the subtitles are read, and on reading comprehension.

2 Background

An extensive amount of qualitative research has been conducted to improve subtitling, most recently due to the introduction of Digital Terrestrial Television (DTT) service. Evaluation of subtitles in the DTT context deals with subtitle style, speed of display, and positioning on the screen.

2.1 Closed Captioning in the U.S.

In the U.S., closed captioning (CC) is governed by the Federal Communications Commission, or FCC, in particular the FCC's Code of Federal Regulations, or CFR 47, Volume 4, Parts 70-79. Specifications for CC style of presentation is detailed in the Closed Captioning Requirements for Digital Television Receivers, as stipulated in FCC Report and Order No. 00-259, detailing the implementation of Section 305 of the Telecommunications Act of 1996, Video Programming Accessibility. The Order adopts the requirement of Section 9 of the Electronic Industries Association standard EIA-708 regarding caption size (large/small), font style (support for eight fonts), color and opacity (eight colors must be implemented by decoders: white, black, red, green, blue, yellow, magenta and cyan), caption location (caption providers are allowed creative control over caption window placement), and multiple caption services (at least one service must be implemented by the decoder). Standard EIA-708-B addresses digital closed captioning implementation (Blanchard, 2003).

2.2 Respeaking in Europe

In Europe, significant progress has been made in the analysis of the quality of audiovisual accessible material, as part of the recently completed Digital Television for All (DTV4All) collaborative project, part of the EC Competitiveness and Innovation Framework Programme (Matamala & Orero, 2010). In particular, as reviewed below for completeness, the project has examined the quality of subtitles' color, position, and display speed.

Dimensions, Encoding, and Transmission. European subtitling standards establish a dimensional limit of 32-37 characters per line of teletext subtitling (when subtitles are transmitted as ASCII characters instead of bitmap images as is the alternative form of subtitle representation found in the DVB sub format, e.g., as used by DVDs) with two lines of text recommended per each subtitle screen (Utray, Ruiz, & Moreiro, 2010).

[Figure 2 about here.]

Font Type and Size. Sans serif fonts are recommended as serifs are considered to be typographical embellishments that tend to impair text legibility as it appears on the screen. Currently, the two top-ranked font rendering styles (see Fig. 2) involve the use of an 80% transparent box or the use of outlined mode. The United Kingdom's Royal National Institute of Blind People specifically recommends the *Tiresias* font for subtitles, although the Spanish Organización Nacional de Ciegos Españoles recommends the *Arial* font, partially due to its wider availability on most computational platforms.

Unlike constant-width fonts, recommendations for point size are usually established empirically by considering the maximum number of pixels provided for a line of text that would fit on the screen, e.g., 576 pixels for SDTV in Europe. The Latin text *Lorem ipsum* is often used for this purpose, suggesting a point size of 31 for *Arial* on PAL standard definition 4:3 screens. The *Lucida Console* font is used by some countries to render teletext subtitles (Univ. Autònoma de Barcelona, 2010).

Font Synchronization and Timing. Subtitles are generally synchronized to appear whenever related characters begin to speak or when sound information is provided, and are usually made to linger on the screen following the six-second rule (Pereira, 2010). Six seconds refers to the amount of time that an average adult hearing or postlocutive deaf viewer (one whose loss of hearing has taken place following development of basic spoken language skills) needs to comprehend the information presented in the two 35-character lines of subtitles.

Subtitle Position. Preference for subtitle position appears to favor their appearance at bottom or in mixed mode but not at top, although such subjective preferences are often influenced by habit or convention (Bartoll & Tejerina, 2010). Further testing of subtitle placement is warranted, particularly in terms of performance (e.g., reading speed and/or comprehension) and process (e.g., eye movement) metrics.

2.3 Eye-Tracking

Compared to the state-of-the-art qualitative research conducted thus far to evaluate subtitles, few inroads have been made into the examination of *process measures*, in the form of eye movements, to corroborate survey responses on which the above research is largely based (Uzquiza, 2010). Early eye-tracking studies indicated that the reading of subtitles is an automatic process that is independent of one's familiarity with subtitling, knowledge of the foreign language in the soundtrack, and the availability of the sound track (d'Ydewalle & Poel, 1999). That is, subtitles are read mandatorily, and processed in detail and remembered. Furthermore, the addition of subtitles (captions) were shown to alter eye movement patterns with the viewing process becoming primarily a reading process although the amount of time spent on subtitles varies among individuals (C. J. Jensema, El Sharkawy, Danturthi, Burch, & Hsu, 2000). As basic reading skills are learned (i.e., first grade reading), more attention shifts to captions since more and more information is obtained from this source (C. Jensema, 2003).

Although eye movements during reading have been studied extensively (Rayner, 1998), thus far, eye movements over video have mainly been evaluated qualitatively, with results largely limited to comparison of scanpath visualizations, reminiscent of very early work, e.g., that of Yarbus (1967). Scanpath visualizations alone do not offer quantitative comparisons of eye movement elements such as fixations, fixation durations, etc. that are commonplace today (Jacob & Karn, 2003; Webb & Renshaw, 2008). According to Uzquiza (2010), this type of research is precisely what is called for to better improve subtitles' processing and reception.

Unfortunately, perusal of the archives of the ACM's Symposium on Eye Tracking Research & Applications reveals a paucity of eye movement analysis over film beyond a handful of notable examples (e.g., (Josephson & Holmes, 2006)). The reason is likely due to the dynamic nature of video: synchronization of video with eye movement data is problematic (e.g., specification of dynamic Areas Of Interest is for the most part absent in commercial software (Papenmeier & Huff, 2010; Ryan, Duchowski, Vincent, & Battisto, 2010)) and the proliferation of digital video formats complicates playback (e.g., commercial systems are often locked in to decoding videos encoded by certain codecs, and transcoding of video can be challenging). A recent study capturing gaze over Japanese *anime*, designed to test "abusive" pop-up gloss found in DVDs, exemplifies the methodological problems of today's commercially-supported eye-tracking analysis of gaze over video (Caffrey, 2009). The study quantitatively showed that a larger percentage of subtitles was skipped in the presence of pop-up gloss, but also highlighted the lack of software available to deal automatically with the large amount of data collected.

Thus far, we are only aware of two efforts that provided quantitative analysis of eye movements over subtitles. The first was conducted under the auspices of the European DTV4All project (Univ. Autònoma de Barcelona, 2010)—tests carried out to date will, on completion, approximate 40,000 subtitles read by hearing, hard of hearing, and deaf participants, constituting the largest corpus of its kind. Presently, pilot data is being evaluated to establish a methodological framework for using eye movement metrics to test various elements of subtitles, including their identification (e.g., color, tag), placement (e.g., top, bottom), justification, background (e.g., box, no box), borders, shadows, emoticons, icons, and speed. The second, and more recent effort, tested the cognitive effectiveness of subtitle processing along with the influence of the subtitles' segmentation quality (Perego, Del Missier, Porta, & Mosconi, 2010). Both studies are reviewed below.

As part of the DTV4All project, results were reported from a study that tested one aspect of chunked respoken subtitles generated for live transmission (Martínez & Linder, 2010). Specifically, the study tested a video clip from the U.K.'s *Six O'Clock News* (airing on 4 July, 2004; see Fig. 3) subtitled by respeaking, where subtitling was either displayed in scrolling mode (word-for-word), or in blocks. Quantitative analysis was performed on the number of fixations made per subtitle line and on the ratio of time spent on subtitles versus the remainder of the scene. Tobii's *Studio* software was used to perform the analysis. The general finding was that viewers (whether hearing, hard-of-hearing, or deaf) devoted about twice as many fixations to scrolling (word-for-word) subtitles as they did to block subtitles. Viewers of scrolling subtitles thus spent a larger proportion of their time (88%) processing text rather than the visual scene, while viewers of blocked subtitles could devote a smaller proportion (67%) of their time to doing the same. Scanpath visualizations of eye movement patterns suggested that fast readers read ahead of scrolling subtitles and cast their gaze (astray fixations) on gaps where they expected to find the next word while slow readers lagged behind and needed to re-read words (regressions).

[Figure 3 about here.]

In line with these results, the more recent study evaluating well- vs. ill-segmented two-line subtitles also found a significantly greater proportion of fixations (more than three times as many) devoted to text than to the visual scene, regardless of the segmentation quality (Perego et al., 2010). Curiously, a larger number of shorter fixations were issued to the subtitle region while fewer longer fixations were devoted to the visual elements. This may potentially suggest a dichotomous "focal/ambient" fixation strategy (Velichkovsky, Joos, Helmert, & Pannasch, 2005) adopted for cognitive processing of text and scene. However, no significant differences were observed in terms of eye movements (numbers of fixations, mean fixation duration, or number of visual transitions between scene/subtitle regions) across conditions of differing subtitle segmentation quality, prompting the authors to conclude that psycholinguistic concerns about subtitle line segmentation are probably overstated.

The present study contributes to prior work by evaluating four subtitling styles:

1. no segmentation (equivalent to blocked subtitling above);
2. word-for-word (equivalent to scrolling above);
3. chunked by phrase (essentially scrolled by phrase); and
4. chunked by sentence (scrolled by sentence).

3 Methodology

This study tests the benefits (if any) of augmenting subtitles with techniques based on the principles of language processing. The method of subtitle segmentation used is somewhat similar to the prior technique based on noun phrase (NP) analysis where NPs (e.g., noun + adjective, noun + prepositional phrase, etc.) were not allowed to break across lines (Perego et al., 2010). To test whether text chunking, i.e., the grouping of a block of text into coherent segments (see Fig. 1), affords subtitle viewing benefit, e.g., in terms of speed of information processing, we developed a custom video playing and eye movement recording application.

The video display program was written in C++ on top of the `ffmpeg` (v0.5) library,[2] an open-source library that contains various codecs facilitating playback of a large (and growing) number of video formats, e.g., H.264, MPEG-2 and MPEG-4, WMV, etc. Of particular interest to library newcomers are the online tutorials by Martin Böhme and Stephen Dranger. The latter's SDL (Simple DirectMedia Layer) tutorial was used to develop a video player using SDL while recording eye movement data delivered by the eye tracker over TCP/IP. Andrew Duchowski's Tobii client library[3] was used to interface with the eye tracker.

While SDL provided system windowing functions, video playback was effected by hardware-accelerated texture mapping of video frames via OpenGL, ensuring appropriate display frame rates while still allowing concurrent recording of eye movement data (performed in a separate thread). To allow subsequent synchronization during visualization playback, the application maintained a global clock to timestamp both eye movement samples (x, y, t) as well as video frames. Although SDL also provides audio playback capability, audio was disabled during the experiment.

[Figure 4 about here.]

[2] `http://www.ffmpeg.org`, last accessed 9/10.

[3] `http://andrewd.ces.clemson.edu/tobii`, last accessed 9/10.

3.1 Stimulus & Means of Subtitle Segmentation

The stimulus video was a short duration (49 s) BBC Breakfast (BBC1) report, airing on 11 July 2009, on the *Gallery on the Green*, an old British Telephone booth (purchased for £1) remodeled to serve as a small art gallery hosting postcard-sized works of art. The booth is located in the Yorkshire Dales. Excerpted video frames are shown in Fig. 4.

The video was subtitled using level 1 teletext for backwards-compatibility with current broadcasting standards (with white font atop a black background box). Subtitles were generated by a live version of the commercial software FAB Subtitler[4] although respeaking was not performed live. The video was first transcribed and then FAB was used to simulate live transmission in each of the four subtitle display modes.

Note that the computational cost of text segmentation is minimal as it depends on the detection of punctuation marks spoken by the respeaker, e.g., "comma" or "full stop", which result in transcribed `NL` and `SEND` commands. As subtitle segmentation is not based on grammatical or pragmatic analysis on the part of the software, it is largely a matter of the human respeaker's performance.

The video (1280 × 720, H.264 MPEG-4) was presented centered on the screen (see below for screen resolution). The visual scene and subtitle regions were delineated by a threshold line set at 683 pixels from the top of the screen (with $(0, 0)$ at top-left), so that gaze points with $y > 683$ were considered falling atop subtitles. Visual attention transitions between regions (termed *saccadic crossovers* below) were defined by successive pairs of gazepoints that crossed this vertical threshold.

3.2 Apparatus

A Tobii ET-1750 video-based corneal reflection eye tracker was used for real-time gaze coordinate measurement. The eye tracker operates at a sampling rate of 50 Hz with an accuracy typically better than 0.3° over a ±20° horizontal and vertical range (Tobii Technology AB, 2003). The eye tracker's 17″ LCD monitor was set to 1280 × 1024 resolution and the stimulus display was maximized to cover the entire screen. The eye tracking server ran on a Windows PC while the client display application ran on a Linux workstation. The client/server PCs were connected via 1 Gb Ethernet.

3.3 Participants

This study involved 28 undergraduate and graduate student participants, 16 male and 12 female. Due to data collection issues, 4 participants' data were discarded from the final analysis, bringing the total down to 24, and dropping the number of males and females to 14 and 10, respectively. The ages of participants ranged from 18 to 47 initially, and 18 to 33 in the final analysis. There were 2 out of the 28 who indicated that English was not their native language, but they were comfortable reading English. These 2 participants were included in the final analysis. In the final analysis, 14 participants reported that they did not wear corrective lenses of any sort, 5 reported that they wore

[4] `http://www.fab-online.com/`, last accessed 9/10.

contacts, and 5 reported that they wore glasses. No participant claimed to have seen the clip before.

[Figure 5 about here.]

3.4 Procedure

Participants sat in front of the eye tracker at a distance of about 60 cm, the tracker camera's focal length (see Fig. 5). Calibration, performed before viewing each video, required visually following nine dot targets displayed sequentially, with each shrinking in diameter from 30 to 2 pixels.

Participants were told that they would be watching four video clips. Before watching the first, they were informed there would be a quiz testing their comprehension of the clip, and were told to pay attention to content in both the scene as well as in the subtitles. Questions on the comprehension quiz were formulated from information found either in the clip's scene or subtitle elements. Participants filled out the comprehension questionnaire before the three subsequent viewings of the same video clip (with differing subtitle styles). They were asked to watch the remaining three clips in the same way as they watched the first.

Participants were also asked to answer subjective questionnaires after each of the four viewings giving their opinions of the subtitles in the clips. After the last clip, participants were asked to fill out a final preference questionnaire to rank the four types of subtitles they had just seen. Rankings were represented by Likert scales, with open-ended questions asking for reasons for a given rating.

3.5 Experimental Design

The study used a completely randomized design with a single factor of subtitle type, varied at four levels. The four subtitling styles were: no segmentation (the area for subtitles was filled with as much text as possible); word by word (words showed up one by one); chunked by phrase (phrases showed up one by one, with one line of the subtitle area being filled at a time); and chunked by sentence (sentences showed up one by one). All clips had audio disabled to ensure information was taken in visually. The video clip itself was held constant across runs; only the subtitles changed.

All participants saw all four videos, but the order in which the videos were played was counterbalanced between participants. In this way, each of the 24 participants saw a unique viewing order. Both qualitative and quantitative data were collected from each run.

Eye tracking data was quantitatively analyzed in terms of the following metrics: mean fixation durations, proportion of gazepoints and fixations in the subtitles, and number of times a viewer's gaze jumped from scene to subtitles, and vice versa (termed *saccadic crossovers*; see Fig. 6). Analysis was performed within- and between-subjects.

[Figure 6 about here.]

[Figure 7 about here.]

4 Results

Eye tracking data were analyzed along the four eye movement metrics, both within-subjects (comparing the four runs shown to each subject) and between-subjects (considering only the data from the first video clip each subject viewed).

With the subtitle segmentation style acting as fixed factor (with subjects serving as the random factor (Baron & Li, 2007)), within-subjects ANOVA showed a significant difference in the number of saccadic crossovers (F(3,68) = 5.60, $p < 0.01$).[5] Pairwise t-tests (no correction) revealed a marginally significant ($p < 0.05$) difference between the word-for-word and chunked by phrase clips. There was also a marginally significant difference ($p < 0.05$) between the word-for-word and chunked by sentence clips (see Fig. 7(a)).

Although between-subjects ANOVA failed to reveal significance of the main effect of subtitle segmentation on the proportion of gazepoints on subtitles (F(3,20) = 1.82, p = 0.18, n.s.) or fixation durations between scene and subtitles (F(3,20) = 1.65, p = 0.21, n.s.), pairwise t-tests (no correction) indicated a marginally significant difference ($p < 0.05$) in fixation durations between the word-for-word and chunked by phrase segmentations. A nearly significant difference (p = 0.06) was observed in the percentage of gazepoints over subtitles between the word-for-word and chunked by phrase segmentations (see Figs. 7(b) and 7(c)).

No differences in preference or comprehension were found.

[Figure 8 about here.]

5 Discussion

The aim of this study was to determine whether text chunking had an effect on the speed with which the subtitles were processed or the overall comprehension of the clip and its subtitles. Although no significant difference was found in comprehension, there were indeed differences in the eye tracking data. Between-subject analysis eliminates the bias of repeated measures by only considering the data collected from the first video each participant viewed, while within-subject analysis allows all data collected to be considered and takes into account the inherent individual differences in eye movement patterns (e.g., idiosyncratic scanpaths (Privitera & Stark, 2000)).

Within-subject analysis revealed a significant difference in the number of saccadic crossovers showing that, on average, an individual tends to switch from scene to subtitle less often than while watching the video with the subtitles chunked by phrase or sentence as compared with word-for-word subtitles. Saccadic crossovers tended to occur while waiting for a word to appear with word-for-word subtitles. Because of the timing at which the words appeared, participants tended to glance up at the scene for a moment before returning to the subtitles in anticipation of the appearance of the next word. The significantly larger number of saccadic crossovers associated with the viewing of the video with word-for-word subtitles was deemed unfavorable compared to the other subtitling methods, notably chunked by phrase, which did not make the viewer

[5] Assuming sphericity as computed by the statistical package R.

wait for words to appear. This allowed for a steadier, more natural viewing experience. These results corroborate previous findings regarding word-for-word subtitling (Univ. Autònoma de Barcelona, 2010) but contradict those where no effect of subtitle segmentation was found (Perego et al., 2010).

A likely reason for the discrepancy in the effect of subtitle segmentation on attentional shifts (our saccadic crossovers) between scene and subtitles is how these shifts are defined. In the previous study, transitions were defined as threshold crossings between consecutive pairs of fixations. In the present situation we defined transitions as threshold crossings between consecutive pairs of gazepoints. Anecdotally, we observed that the fixational filter (set to a minimum duration of 100 ms and 30 pixel radius as in the previous study) is overly aggressive in removing fast gaze transitions, which we believe occur with greater frequency when viewing video than when viewing still images, the type of stimulus for which the filter was probably originally designed. We thus turned off fixational filtering in our analysis and used the term *saccadic crossover* to emphasize this (perhaps subtle yet critical) point.

Between-subject analysis indicated that participants tended to devote a smaller percentage of gazepoints to subtitles when chunked by phrase. In contrast, word-for-word subtitling appeared to elicit the largest percentage of gazepoints. The percentage of gazepoints devoted to subtitles parallels the amount of time that participants spent looking at the subtitles. This observation favors the chunked by phrase subtitles, as they required the least amount of viewing time.

Between-subject analysis also showed a trend toward longer mean fixation durations over the word-for-word subtitles, when compared to the chunked by phrase subtitles. Fixation duration corresponds to cognitive processing of the information that is being fixated. Just and Carpenter (1980) posited that the eye-mind assumption holds for reading text; that is, the length of a fixation on a word is proportionate to the amount of time that the reader spends processing it. If so, subtitles chunked word-for-word tend to take longer to process than subtitles chunked by phrase.

In the absence of significant differences in preference or comprehension, eye tracking provides a compelling means of inferential analysis as the subtitles' effect is seen solely in the eye movement patterns they elicit. Subtitles chunked by phrase appeared to provide the best relative viewing experience, as they evoked the least number of crossovers and the smallest percentage of gazepoints. Conversely, subtitles chunked word-for-word appeared to provide the relative worst viewing experience, as they evoked the largest number of crossovers and the largest percentage of gazepoints.

If mean fixation duration is representative of the speed with which information is processed, the data imply that subtitles chunked by phrase were the easiest to process, and those chunked word-for-word were the most difficult. Subtitles chunked by phrase should detract the least from one's viewing experience with no detrimental effect on comprehension.

Aggregate heatmap visualizations of all participants' gaze atop each subtitling style is shown graphically in Fig. 8. Heatmaps project temporal gaze data onto a single (arbitrary) frame, qualitatively showing that, over time, gaze is distributed over subtitles with lower frequency when subtitles are chunked by phrase or by sentence.

6 Study Limitations

Hearing participants' reception of written text differs from that of the deaf or hard-of-hearing (Cabeza-Pereiro, 2010), whose first language may have no written form. Text chunking impacts the way subtitles were viewed by hearing participants, but further research is needed to better understand the effect for both hearing as well as the Deaf and Hard-Of-Hearing.

7 Conclusion

Examination of hearing participants' eye movements over subtitled video showed that different styles of text segmentation elicit different viewing behaviors. According to eye movement metrics, text chunking by phrase or by sentence reduces the amount of time spent on subtitles, and presents the text in a way that is more easily processed. Qualitative and quantitative studies such as these are prompting discussion of adoption of this form of live subtitling by European subtitling broadcast services and media companies. In the U.S., although President Obama signed the 21st Century Communications & Video Accessibility Act into law on 8 October 2010, it is not yet clear what its effects on closed captioning will be.

Acknowledgments

This work was supported in part by CNS grant #0850695 from the National Science Foundation (REU Site: Undergraduate Research in Human-Centered Computing) and by Spanish Ministry of Science and Innovation grant FFI2009-08027 and with Catalan Government funds 2009SGR700.

References

Baddeley, A. (2003). Working memory and language: an overview. *Journal of Communication Disorders*, *36*, 189–208.

Badre, A. N. (1982, July/August). Selecting and representing information structures for visual presentation. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-12*(4), 495–504.

Baron, J., & Li, Y. (2007, 09 November). *Notes on the use of R for psychology experiments and questionnaires.* Online Notes. (URL: `http://www.psych.upenn.edu/~baron/rpsych/rpsych.html` (last accessed Dec. 2007))

Bartoll, E., & Tejerina, A. M. (2010). The positioning of subtitles for the deaf and hard of hearing. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 69–86). Bern, Switzerland: Peter Lang AG.

Blanchard, R. N. (2003, August). EIA-708-B Digital Closed Captioning Implementation. *IEEE Transactions on Consumer Electronics*, *49*(3), 567–570.

Boulianne, G., Beaumont, c., Jean-Fran% , Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., et al. (2010). Shadow speaking for real-time closed-captioning of TV broadcasts in French. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 191–207). Bern, Switzerland: Peter Lang AG.

Buxton, W. (1986). Chunking and phrasing and the design of human-computer dialogues. In *Proceedings of IFIP World Computer Congress* (pp. 475–480).

Cabeza-Pereiro, C. (2010). The development of writing and grammatisation: The case of the deaf. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 45–57). Bern, Switzerland: Peter Lang AG.

Caffrey, C. (2009). *Relevant abuse? Investigating the effects of an abusive subtitling procedure on the perception of TV anime using eye tracker and questionnaire*. Unpublished doctoral dissertation, Dublin City University, Dublin, Ireland. (`http://doras.dcu.ie/14835/1/Colm_PhDCorrections.pdf` (last accessed Sep. 2010))

Chafe, W. L. (1980). The Deployment of Consciousness in the Production of a Narrative. In W. L. Chafe (Ed.), *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Norwood, NJ: Ablex Publishing Corporation.

d'Ydewalle, G., & Poel, M. Van de. (1999). Incidental Foreign-Language Acquisition by Children Watching Subtitled Television Programs. *Journal of Psycholinguistic Research*, *28*(3), 227–244.

Faraday, P., & Sutcliffe, A. (1999). Authoring Animated Web Pages Using Contact Points. In *Human Factors in Computing Systems: CHI 99 Conference Proceedings* (pp. 458–465). ACM Press.

Jacob, R. J. K., & Karn, K. S. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573–605). Amsterdam, The Netherlands: Elsevier Science.

Jensema, C. (2003, September). *The Relation Between Eye Movement and Reading Captions and Print by School-Age Deaf Children* (Final Report No. Grant Award Number HH327H000002). Wheaton, MD: Institute for Disability Research and Training, Inc.

Jensema, C. J., El Sharkawy, S., Danturthi, R. S., Burch, R., & Hsu, D. (2000, July). Eye Movement Patterns of Captioned Television Viewers. *American Annals of the Deaf*, *145*(3), 275–285.

Josephson, S., & Holmes, M. E. (2006). Clutter or Content? How On-Screen Enhancements Affect How TV Viewers Scan and What They Learn. In *Eye Tracking Research & Applications (ETRA) Symposium* (pp. 155–162). San Diego, CA.

Just, M. A., & Carpenter, P. A. (1980, July). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Lambert, S. (1989). Simultaneous interpreters: One ear may be better than two. *TTR: traduction, terminologie, rédaction*, *2*(1), 153–162.

Martínez, J., & Linder, G. (2010, October 6-8). The Reception of a New Display

Mode in Live Subtitling. In *Proceedings of the Eighth Languages & The Media Conference* (pp. 35–37).

Matamala, A., & Orero, P. (Eds.). (2010). *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing*. Bern, Switzerland: Peter Lang AG.

Miller, G. A. (1956, March). The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity For Processing Information. *Psychological Review*, *63*(2), 81–97.

Papenmeier, F., & Huff, M. (2010). DynAOI: A tool for matching eye-movement data with dynamic areas of interest in animations and movies. *Behavior Research Methods*, *42*(1), 179–187.

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The Cognitive Effectiveness of Subtitle Processing. *Media Psychology*, *13*(3), 243–272. (URL: `http://dx.doi.org/10.1080/15213269.2010.502873` (last accessed Sep. 2010))

Pereira, A. (2010). Criteria for elaborating subtitles for deaf and hard of hearing adults in Spain: Description of a case study. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 87–102). Bern, Switzerland: Peter Lang AG.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *22*(9), 970–982.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422.

Ribas, M. A., & Romero-Fresco, P. (2008). A Practical Proposal for the Training of Respeakers 1. *The Journal of Specialised Translation*, *10*. (URL: `http://www.jostrans.org/issue10/art_arumi.php`, (last accessed Sep. 2010))

Ryan, W. J., Duchowski, A. T., Vincent, E. A., & Battisto, D. (2010, March 22-24). Match-Moving for Area-Based Analysis of Eye Movements in Natural Tasks. In *Eye Tracking Research & Applications (ETRA)*. Austin, TX: ACM.

Tobii Technology AB. (2003). *Tobii ET-17 Eye-tracker Product Description.* ((Version 1.1))

Univ. Autònoma de Barcelona. (2010, February). *Digital Television For All (DTV4All)* (Tech. Rep.). University of Roehampton, UK. (URL: `http://dea.brunel.ac.uk/dtv4all/ICT-PSP-224994-D25.pdf`, (last accessed Sep. 2010))

Utray, F., Ruiz, B., & Moreiro, J. A. (2010). Maximum font size for subtitles in Standard Definition Digital Television: Test for a font magnifying application. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 59–68). Bern, Switzerland: Peter Lang AG.

Uzquiza, V. A. (2010). SUBSORDIG: The need for a deep analysis of data. In A. Matamala & P. Orero (Eds.), *Listening to Subtitles: Subtitles for the Deaf and Hard of Hearing* (pp. 163–174). Bern, Switzerland: Peter Lang AG.

Velichkovsky, B. M., Joos, M., Helmert, J. R., & Pannasch, S. (2005, 21-23 July). Two Visual Systems and their Eye Movements: Evidence from Static and Dynamic Scene Perception. In *CogSci 2005: Proceedings of the XXVII Conference of the Cognitive Science Society* (pp. 2283–2288).

Webb, N., & Renshaw, T. (2008). Eyetracking in HCI. In P. Cairns & A. L. Cox (Eds.), *Research Methods for Human-Computer Interaction* (pp. 35–69). Cambridge, UK: Cambridge University Press.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.

List of Figures

1 Examples of subtitles displayed with and without punctuation-based segmentation. 16

2 Top two qualitative rankings of subtitle styles, with (a) finding broadest acceptance and composed of sans-serif font displayed in a 80% transparent window, and (b) ranking second, composed of sans-serif font displayed in outlined mode (Univ. Autònoma de Barcelona, 2010). . . 17

3 Scrolled subtitles (a) require more time reading than blocked subtitles (b). An example eye movement visualization shows a reader of scrolled subtitles casting their gaze on the word *patients*, then regressing to the previous word *several*, before finally regressing to *we've got*. In contrast, a reader of blocked subtitles reads the same line of text by issuing only four sequential fixations to *we've*, *several*, *patients*, and *that* to acquire the line's meaning (Univ. Autònoma de Barcelona, 2010). . . 18

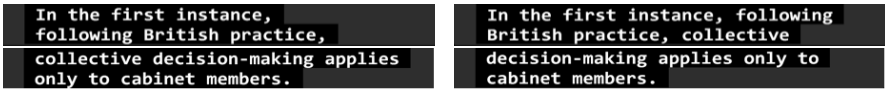
4 Frames excerpted from video used as stimulus. 19

5 Example of participant at eye tracker during calibration. During calibration, a yellow dot moves to nine fixation targets while two grey dots display the participant's eyes in the camera's reference frame (as an indication of their relative position to the camera). 20

6 A *saccadic crossover* is recorded when a gazepoint in the subtitles is immediately followed by a gazepoint in the scene, or vice versa. . . . 21

7 Within- and between-subjects results. 22

8 Heatmap visualizations showing accumulated gaze from all viewers on a representative frame of video with different subtitle display styles. . 23



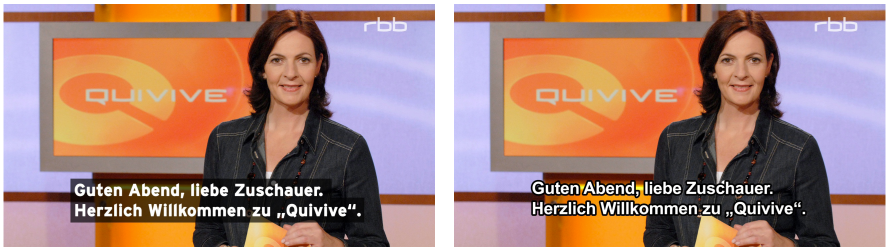
In the first instance,
following British practice,
collective decision-making applies
only to cabinet members.

(a) With punctuation-based segmentation.

In the first instance, following
British practice, collective
decision-making applies only to
cabinet members.

(b) Without punctuation-based segmentation.

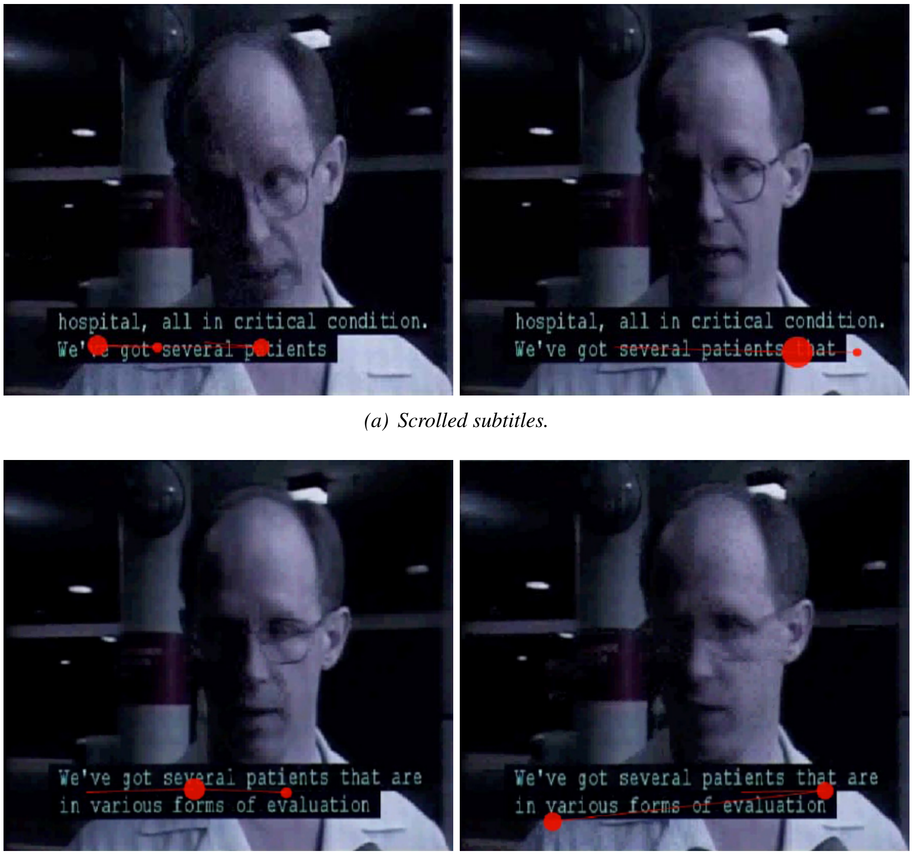
Figure 1: Examples of subtitles displayed with and without punctuation-based segmentation.


QUIVIVE
rbb
Guten Abend, liebe Zuschauer.
Herzlich Willkommen zu „Quivive“.
QUIVIVE
rbb
Guten Abend, liebe Zuschauer.
Herzlich Willkommen zu „Quivive“.

(a) Best-ranked subtitle style.

(b) Second-ranked subtitle style.

Figure 2: Top two qualitative rankings of subtitle styles, with (a) finding broadest acceptance and composed of sans-serif font displayed in a 80% transparent window, and (b) ranking second, composed of sans-serif font displayed in outlined mode (Univ. Autònoma de Barcelona, 2010).


hospital, all in critical condition.
We've got several patients
hospital, all in critical condition.
We've got several patients that
We've got several patients that are
in various forms of evaluation
We've got several patients that are
in various forms of evaluation

(a) Scrolled subtitles.

(b) Blocked subtitles.

Figure 3: Scrolled subtitles (a) require more time reading than blocked subtitles (b). An example eye movement visualization shows a reader of scrolled subtitles casting their gaze on the word *patients*, then regressing to the previous word *several*, before finally regressing to *we've got*. In contrast, a reader of blocked subtitles reads the same line of text by issuing only four sequential fixations to *we've*, *several*, *patients*, and *that* to acquire the line's meaning (Univ. Autònoma de Barcelona, 2010).

BBC NEWS 07:46
POST OFFICE
a new name joins the list. Up in the
Yorkshire Dales, you will find the
BBC NEWS 07:46
say the idea is just a little crazy.

Figure 4: Frames excerpted from video used as stimulus.

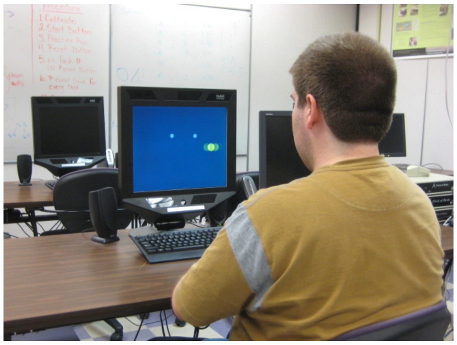


Figure 5: Example of participant at eye tracker during calibration. During calibration, a yellow dot moves to nine fixation targets while two grey dots display the participant's eyes in the camera's reference frame (as an indication of their relative position to the camera).

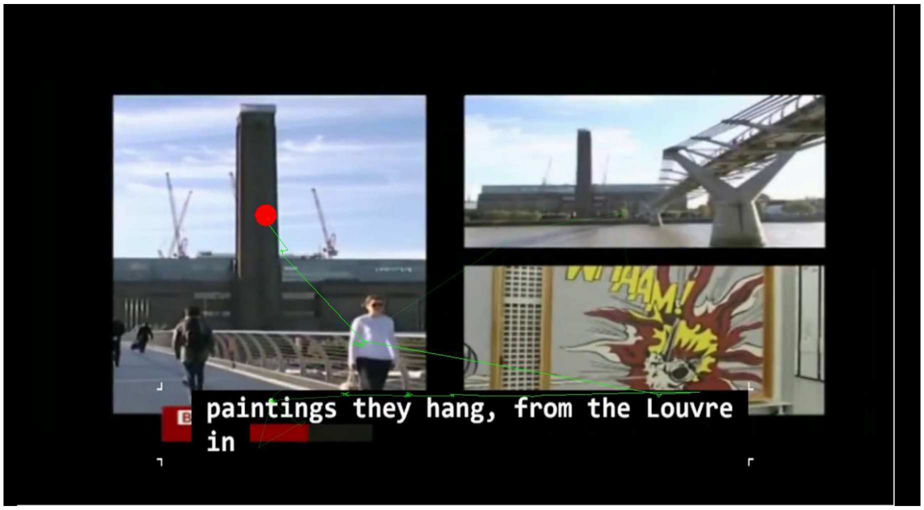
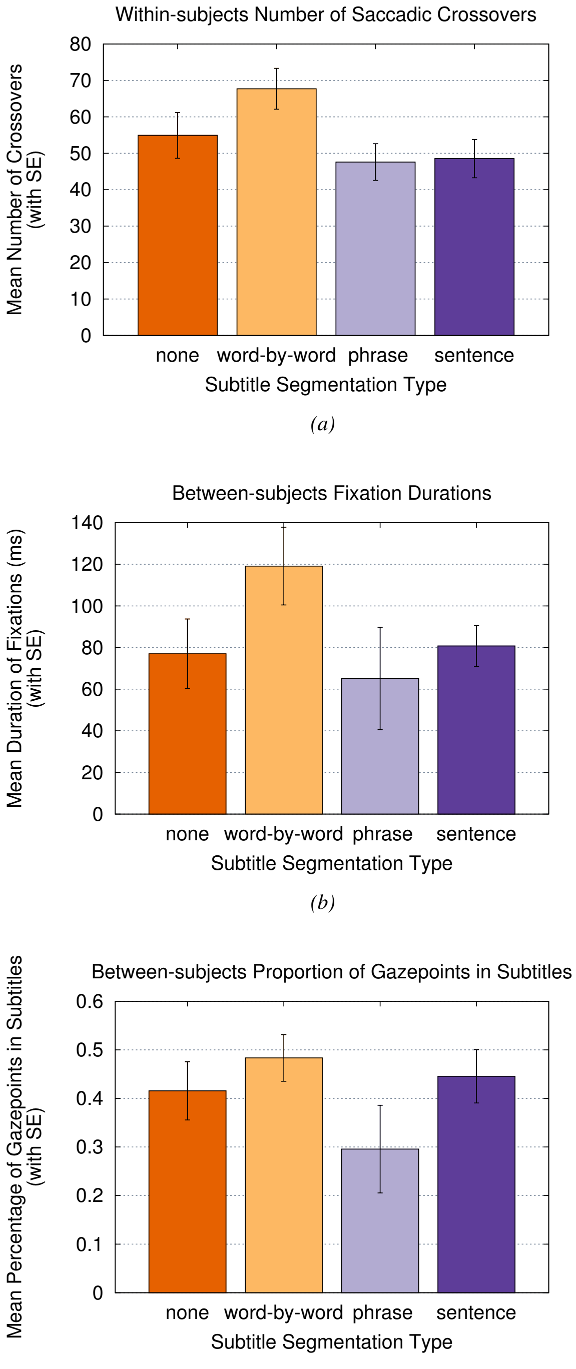


Figure 6: A *saccadic crossover* is recorded when a gazepoint in the subtitles is immediately followed by a gazepoint in the scene, or vice versa.

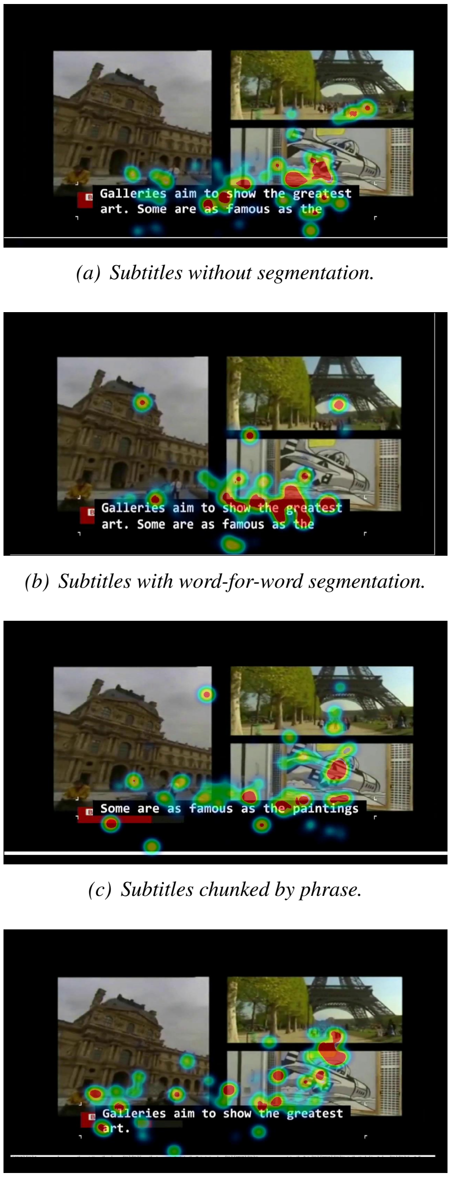


(a)

(b)

(c)

Figure 7: Within- and between-subjects results.


Galleries aim to show the greatest
art. Some are as famous as the
Galleries aim to show the greatest
art. Some are as famous as the
Some are as famous as the paintings
Galleries aim to show the greatest
art.

(a) Subtitles without segmentation.

(b) Subtitles with word-for-word segmentation.

(c) Subtitles chunked by phrase.

(d) Subtitles chunked by sentence.

Figure 8: Heatmap visualizations showing accumulated gaze from all viewers on a representative frame of video with different subtitle display styles.