www.thecine-files.com

THE CINE-FILES

DOSSIER ON FILM SOUND

Listening to Images: Audio Description, the Translation Overlay, and Image Retrieval

Mara Mills

Link to film clip in criticalcommons.org: "Audio Described Version of *The Unconquered* (1953). Clip courtesy of The Described and Captioned Media Program."

In *Picture Theory* (1994), W.J.T. Mitchell expands upon the ancient art of *ekphrasis*—the verbal description of visual material. Poems about paintings, radio shows about photographs, and other "verbal-visual encounters" (he writes) take several forms. Texts conjure, ventriloquize, repress, and objectify images. Some images are themselves linguistic. Mitchell contrasts the "purely figurative" ekphrastic event with "the encounters of verbal and visual representation in 'mixed arts' such as illustrated books, slide lectures, theatrical presentations, film, and shaped poetry." [1] For those cases, he offers a set of typographic distinctions to clarify the nature of the relationship between the visual and the verbal:

> The slash to designate '"image/text" as a problematic gap, cleavage, or rupture in representation. The term "imagetext" designates composite, synthetic works (or concepts) that combine image and text. "Image-text," with a hyphen, designates *relations* of the visual and verbal. [2]

Over the past two decades, multimedia encounters have proliferated as a result of new digital tools, putting pressure on literary and visual theories of interpretation, perception, and translation. Semioticians continue to elaborate taxonomies of audio, audiovisual, and interactive signs, while media scholars specify numerous interrelations between sound and image. [3] The See This Sound project—an exhibition and web archive hosted by the Academy of Visual Arts Leipzig—explores audio-visual relations that manifest as conversion (e.g. sonification, graphic notation); representation (e.g. synchronization, audiovisual montage); and perception (e.g. synesthesia, sensory integration). The field of translation studies is similarly overgrown with source media (video games, museum tours, bilingual conversations) and target audiences (blind, D/deaf, intercultural). Linguistic translation has been transformed by machine automation as well as new platforms for multimodal communication. In the professional subfield of "audiovisual translation," images and sounds are converted into verbal form through subtitling (the addition of writing to a medium: closed captioning, bilingual or intralingual subtitling, surtitling, intertitling) and through revoicing (the addition of voice: lip-sync dubbing, voice-over, free commentary, interpreting, audio description). [4] Most of these examples simply entail new forms of display for verbal translation. However, audio description—especially applied to moving images—extends translation across words, modes, and media.

Jay Dolmage suggests that we think of audio description—the translation of visual and audiovisual media into words for blind spectators—as a genre of ekphrasis. In *Disability Rhetoric*, Dolmage discusses the destigmatizing effect of this move; ekphrasis encourages us to imagine "'accommodations' for people with disabilities as adding artistic and rhetorical value, not simply transposing or distilling meanings." [5] In the United States, the technique of audio description began to be formalized in the 1970s. Today it is employed to provide access to performances, photographs, moving images, and exhibits: live theater, graphic novels, historical landmarks, art shows, television, sound film, webpages. In the clip that opens this essay, taken from the audio described version of the 1953 documentary *The Unconquered*, description is interpolated into the pauses of the film's narration. Katharine Cornell's voice-over otherwise emphasizes Helen Keller's communicative alterity while threading together the disparate elements of the film—photographs, newsreel footage, segments from the 1919 biopic *Deliverance,* and new shots. The audio description is at odds with the narration and the musical theme, conveying Keller's activity in a wholly unsentimental manner.

Scholars have enlarged the concept of ekphrasis to encompass the description of any medium by another, from musical portrayals of paintings to literary renditions of dances. In practice, audio description—which also goes by the terms "video description," "verbal description," "visual description," and simply "description"—often exceeds visual-verbal translation. Depending on the source medium, it can include the reading aloud of text; explanatory remarks on sound cues, noises, and musical themes; and descriptive narration about visual elements such as settings, actions, costumes, and facial expressions. It can take the form of amateur or professional live narration (e.g., at theater events), pre-recording (e.g., for gallery tours), or recording and overdubbing (e.g., for videos). Guidelines for audio description vary somewhat by agency, especially regarding the description of race, ethnicity, sexuality, emotion, sexually explicit material, and humor. Some translators recommend the description of form—for instance, cinematic editing strategies—as well as content, taking into consideration viewers who are partially sighted or who became blind later in life. [6] In 2010, the Royal National Institute of Blind People (RNIB) published a comparative report on audio description standards in six countries: the U.K., U.S., France, Germany, Greece, and Spain. The report noted extreme constraints when translating an audiovisual moving image into words:

> The common denominator across all the guidelines is that description should only be added during pauses in a film/TV programme and at no cost should the description undermine the film/television programme. In conjunction with the advice in the German standards, A description should really only follow when the film is completely silent, so when there is no dialogue or noise or music. However, this hardly ever happens and therefore one has to make a decision to speak over music and also some noises. In doing this, one has to continually question whether the precise place in which one wants to talk over the audio fulfills an important function in terms of the mood and atmosphere and thus whether it should remain undisturbed. Music and sounds are also part of the language of a film! [7]

Unlike ekphrasis, audio description thus often merges with and transforms the thing-described. In this brief essay, I want to propose that audio description is part of a growing category of media use, the "translation overlay," in which alternative content is added to source material without creating a new work. By "alternative content," a phrase I take from the Web Access Initiative (W3C), I refer to the broad ambit of contemporary translation studies: linguistic translation, sensory modality translation, transcription, and revoicing. Translation overlays include captioning, fansubbing, fandubbing, scanlation, embedded sign language translation, karaoke, voice-over, ADR (automated dialogue replacement), bilingual editions, and "twin-vision" braille/print books. [8] Arguably, the concept also includes interpretive addenda to a soundtrack such as overdubbed Foley and clean audio. Translation overlays can be amateur or professional. The phenomenon traverses disability, language translation, intercultural performance, and sound design.

The translation overlay can be distinguished from other forms of translation output—such as text-to-speech screen reading—because it integrates a new track into the original work. The translation overlay is a supplement to other theories of media use, such as remix and reenactment, in its dual emphasis on formal transformation and semiotic homology. Scholars who study anime and manga have noted the recalcitrance of fansubs and scanlation (amateur subtitles or translations) to current media theory. Cultural anthropologist Mimi Ito and legal theorist Jordan Hatcher have argued that fansubs and scanlations aren't accounted for by the concept of the remix; as Hatcher writes, "their aim is to remain faithful to the original work." [9] Similarly Jeremy Douglass, Lev Manovich, and William Huber claim:

> The creative activity of scanlation groups is neither "authorship" nor "remix." It also cannot be adequately described using a well-known distinction by Michel de Certeau between "strategies" and "tactics" (because in contrast to the unconscious tactics described by de Certeau, scanlation groups add new pages to manga series they publish quite consciously.) Similarly, scanlations are neither "remediations" (Jay David Bolter and Richard Grusin) nor "transmedia" (Henry Jenkins). In short, we currently lack proper terms to describe them. [10]

Douglass, Manovich, and Huber recommend that scanlations be classified simply as "versions." However, this does not capture the double weave of source and target text. The portmanteaus of scanlation and fansubbing run deeper than wordplay.

Translation overlay has become increasingly precise and convenient with the availability of multitrack recording and digital compositing. In the realm of digital media production, overlay is one among many compositing techniques. Overdubbing is the parallel term for layering in a new audio track. [11] In *The Language of New Media*, Manovich contrasts compositing with the more disjointed aesthetic of montage. Compositing is a central operation across new media; it "exemplifies a more general operation of computer culture—assembling together a number of elements to create a singular seamless object." [12] Compositing can be a mode of primary production as well as remixing. Similarly, the overlay function has diverse applications, from subtitling to watermarking to branding to transfiguring an image. The translation overlay is distinguished by the melding of a work with its own translation, and it exceeds the use of digital tools.

If audio description offers an example of the "translation overlay" genre of media use, it also points to an emerging field of applications for audiovisual translation that depend upon description. Audio described videos have been advertised as aids for multimedia literacy; for "eyes-free viewing" by sighted people (e.g. while driving); for video-based medical education; and for training Autistic people to read facial expressions. Moreover audio description played a significant role in the early history of efforts to index, catalog, search, and retrieve digital images. In the 1990s, a group of film archivists and library scientists, alongside licensing firms like Corbis, began to research audio description, in the form of transcripts or speech-to-text, as a means to expedite and even automate image indexing. Otherwise, archivists and art librarians assigned keywords to images manually via a number of different indexing schemes. [13] Before 2000, when the addition of metadata to digital images was scarce and Google Images had not yet launched, researchers hoped that pre-existing caches of image descriptions could be used to generate keyword or caption tags. [14] Audio description was also a resource for understanding the ways people classified and searched for images (their "visual information-seeking behavior").

When web image search came to rely on user-entered text such as filenames and weblinks, information scientists continued to investigate audio description—as well as closed captioning—as means to automatically assign text to the individual shots or scenes in moving images. This "shot-by-shot indexing" promised to dramatically improve the precision of search and retrieval in moving image databases. As one report explains, "Since shot-level indexing of moving images is very expensive, it is attractive to use sources that already exist in electronic and textual form." [15]

Information scientists now pursue content-based image retrieval—searching images themselves as opposed to their metadata—through "machine translation from images to text." [16] Computers recognize features in images and generate labels for their contents, making these images retrievable by users searching via text query. This procedure has recently been demonstrated in several machine learning experiments based on massive image annotation datasets, supplied by Mechanical Turk workers or by sources such as Flickr. According to one team of researchers:

> Mining the absolutely enormous amounts of visually descriptive text available in special library collections and on the web in general, makes it possible to discover statistical models for what modifiers people use to describe objects and what prepositional phrases are used to describe relationships between objects. These can be used to select and train computer vision algorithms to recognize constructs in images. [17]

Along with web image search, content-based image retrieval is projected to transform medical diagnosis, robot navigation, and—through the recognition of actions captured on CCTV—surveillance. [18]

All of this suggests the growing pervasiveness of description as a mode of human and machine translation. Ekphrasis, while by no means the only verbovisual or audiovisual relation, persists in the domain of new media—streamlined and operationalized. "Sentences are rich, compact, and subtle representations of information," one image-retrieval researcher writes, "Even so, we can predict good sentences for images that people like." [19] Analyzing translation overlays, such as audio description and closed captioning, is one route to thinking critically about what counts as "a good sentence for an image"—and about the statistically-derived translations that increasingly underlay images, embedded in databases.

Audio description adds another track to film sound. It signals a broader genre of "translation overlay" in media production, and it demands new theories of audiovisual translation—theories that account for the role of description, including machine annotation, in the move from visual to verbal.

Mara Mills is an Assistant Professor of Media, Culture, and Communication at New York University, working at the intersection of disability studies and media studies. She is currently completing a book titled *On the Phone: Deafness and Communication Engineering*. Articles from this project can be found in *Social Text*, *differences*, the *IEEE Annals of the History of Computing*, and *The Oxford Handbook of Sound Studies*. Her second book project, *Print Disability and New Reading Formats*, examines the reformatting of print over the course of the past century by blind and other print disabled readers, with a focus on Talking Books and electronic reading machines.

Notes

[1] W.J.T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation* (Chicago: University of Chicago Press, 1994), 157.

[2] Ibid., 89.

[3] See, for instance, Winfried Nöth, ed., *Semiotics of the Media: State of the Art, Projects, and Perspectives* (Berlin: Walter de Gruyter & Co., 1997); Michel Chion, *Audio-Vision: Sound on Screen*, trans. Claudia Gorbman (New York: Columbia University Press, 1994).

[4] For a more detailed overview of audiovisual translation (sometimes called multimedia translation), see Mona Baker and Gabriela Saldanha, eds., *Routledge Encyclopedia of Translation Studies*, 2nd edition (New York: Routledge, 2009). Audiovisual translation is distinguished from the related terms "adaptation" and "intermedial transposition." On adaptation, see pp. 5-8 of the *Encyclopedia*.

[5] Jay Dolmage, *Disability Rhetoric* (Syracuse: Syracuse University Press, 2014), 140. Susan Schweik made a similar argument in a June 2012 workshop held at the University of California Humanities Research Institute, attended by the author. This workshop is described in Catherine Kudlick and Susan Schweik, "Collision and Collusion: Artists, Academics, and Activists in Dialogue with the University of California and Critical Disability Studies," *Disability Studies Quarterly* 34, no. 2 (2014), http://dsq-sds.org/article/view/4251/3609.

[6] Louise Fryer and Jonathan Freeman, "Cinematic Language and the Description of Film: Keeping AD Users in the Frame," *Perspectives: Studies in Translatology* 21, no. 3 (2013): 412-426.

[7] "A Comparative Study of Audio Description Guidelines Prevalent in Different Countries," *Media and Culture Department, Royal National Institute of Blind People,* December 15, 2010, http://audiodescription.co.uk/uploads/general/RNIB._AD_standards.pdf

[8] Karaoke subtitling frequently appears in handbooks of translation studies; arguably karaoke singing is another form of translation overlay. Interlingual subtitling often finds mention alongside modality translation techniques. See, for instance, the W3C guidelines for media accessibility or Carmen Millán and Francesca Bartrina eds., *The Routledge Handbook of Translation Studies* (New York: Routledge, 2013).

[9] Jordan Hatcher, "Of Otakus and Fansubs: A Critical Look at Anime Online in Light of Current Issues in Copyright Law," *SCRIPT-ed* 2, no. 4 (2005): 517.

[10] "Understanding Scanlation: How to Read One Million Fan-Translated Manga Pages," *Image & Narrative* 12, no. 1 (2011): 193.

[11] The audio description of moving images is called both "audio overlay" and overdubbing. Florian Grond has suggested the phrase "translation inlay" to me, which more clearly conveys the amalgamation of translation and original. I've chosen "overlay" as a familiar term from editing and compositing.

[12] Lev Manovich, *The Language of New Media* (Cambridge, MA: The MIT Press, 2001), 143.

[13] This is the subject of an in-progress dissertation by Diana Kamin, Ph.D. candidate in the Department of Media, Culture, and Communication at New York University.

[14] The term "caption" here refers to a description rather than a subtitle translation. For history and state-of-the field circa 2000, see Abby Goodrum, "Image Information Retrieval: An Overview of Current Research," *Informing Science* 3, no. 2 (2000): 63-67.

[15] Subtitles were also used. Alt.text was less useful at the time because that field tended to contain a title rather than an image description. Moreover it was often left blank. Ernie Dornfeld, "Classification and Indexing for Image Collections: Theory and Practice," *Bulletin of the American Society for Information Science* (December/January 1998): 15.

[16] Girish Kulkarni et al., "Baby Talk: Understanding and Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 12 (December 2013): 2898.

[17] Ibid., 2891.

[18] For an early overview of the field and its applications, see John P. Eakins and Margaret E. Graham, *Content-Based Image Retrieval, A Report to the JISC Technology Application Programme. Technical report* (Newcastle: Institute for Image Data Research, University of Northumbria, January 1999). For an updated and accessible introduction to the field, see: John Markoff, "Researchers Announce Advance in Image-Recognition Software," *New York Times*, 17 November 2014. With thanks to Solon Barocas for pointing me to this article. Note: other image retrieval systems, such as those used for facial recognition, rely on feature matching rather than text queries.

[19] Eli Farhadi et al., "Every Picture Tells a Story: Generating Sentences from Images," *Computer Vision—ECCV 2010, Part IV* (2010): 12. See also Olga Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," ArXiV1409.0575v3 (2015), http://arxiv.org/abs/1409.0575. Thanks to Lev Manovich for forwarding this article to me, and for his comments on this essay.