Sada Vehra: A Framework for Crowdsourcing Punjabi Language Content

Jasmine Hentschel
University of Michigan School of Information
105 S. State Street
Ann Arbor, MI 48109
hentscj@umich.edu

Joyojeet Pal
University of Michigan School of Information
105 S. State Street
Ann Arbor, MI 48109
joyojeet@umich.edu

ABSTRACT
We present preliminary research leading to a prototype for Sada Vehra, an online resource for crowdsourcing Punjabi language content. We argue that languages less represented online benefit from crowdsourced volunteer-contributed translation, and present some challenges around effective crowdsourced non-expert translations. These include interface challenges around handoffs and ownership, language-specific challenges such as nuance and standardization, and infrastructure-specific challenges such as bandwidth and media quality.

Categories and Subject Descriptors
K.4.3 [**Organizational Impacts**]: computer-supported collaborative work

General Terms
Design, Human Factors

Keywords
Crowdsourcing, Translation, ICTD, HCI, Punjabi

1. INTRODUCTION
The representation of languages from the Global South online has been a fundamental challenge within the ICTD world, impacting the access individuals from less represented language groups have to digital information. While there has been a growth of East Asian language content online, the continuing dominance of English, and to some extent Spanish and French, has meant that access to content has often been restricted to elites in postcolonial states [5]. However, the presence of bilingual populations online also presents the opportunity to preserve and promote linguistic resources in new ways.

ICTD '15, May 15 - 18, 2015, Singapore, Singapore Copyright 2015 ACM 978-1-4503-3163-0/15/05◄$15.00
http://dx.doi.org/10.1145/2737856.2737901

Crowdsourcing translations is a practice gaining traction and completely shifting the translation industry to one that is participatory and open to non-professionals in ways different than ever before. Alongside the rise of machine translation that works with variable efficiency in different language groups, several online platforms have focused on increasing the corpus of translated material by promoting volunteer effort among individuals with personal connection to or interest in content rather than solely financial incentives. Variations of processes for crowdsourcing translations in recent years have been used for such purposes as education, disaster relief management, citizen journalism, and entertainment [4].

Punjabi is spoken by nearly 100 million people, [7,8] with mother tongue speakers primarily residing in India and Pakistan. Despite the large number of speakers and vast diaspora, the language has a very limited web presence, likely because a great number of Punjabi speakers online tend to be bilingual in English [14].

The primary goal is to increase access to Punjabi-language content as a cultural resource. We propose this by creating an integrated workflow and resource portal that supports the creation of English translations for Punjabi-language multimedia content. The portal is based on the principle that bilingual speakers can contribute on a sliding scale based on their language skills and time availability with the right scaffolding and access to resources. Translation and subtitling extend a body of cultural content, but also help in sharpening the language skills of the translators themselves.

2. RELATED WORK
Although there has been much work on access to technological resources by people from underrepresented languages within ICTD in the last few years, there is very little work on building resources to preserve language and cultural materials. Two bodies of related work are immediately relevant—the work around motivating crowdsourced volunteerism and that around workflow and resources for translation projects.

2.1 Encouraging Participation
Motivations for participating in crowdsourcing typically come from a combination of factors—some intrinsic, such as challenge, fun, or enjoyment—and some extrinsic, like wanting to gain new skills, participate in a community, share culture, or engage with others [3]. Likewise, the research shows that participative systems can be built by compelling members to care about what peers think of their contributions and making them feel that their

participation matters [ibid]. Breaking down systems into parts and encouraging participation by users with a range of language skills is conducive to greater participation [2]. A motivating factor for participating in translation activities among language learners is that it's widely seen as an important strategy in second language acquisition [13]. While these aspects contribute to bringing volunteers into participating in the system, altruism and a desire to see the strengthening of one's cultural resources are often central to peoples' motivation and willingness to continue contributing.

2.2 Workflow Considerations

Audiovisual translation until recently has centered primarily on film subtitling and dubbing. Within the subtitling industry, procedures and methods vary widely depending on the organization doing the work and requirements of the project [16]. This could be in part because of the myriad ways professionals acquire their skills, which are "often driven by prescriptivist judgments and not always based on research" [15].

When employing crowds of non-professionals to do translations, clearly defined tasks and articulated piece-meal goals that reduce control of flow or content to the crowd are important [3]. An important workflow consideration is building redundancy into the system. Accuracy of translations can be hard to measure and using non-professionals to complete a cognitively demanding task like translation requires a certain degree of quality control. In the broadest sense, it can be exercised through a segmented workflow that includes at least one round of revision and review, a standard practice in the translation industry [10].

2.3 Translator Resources and Tools

Professionals and non-professionals alike employ a wide range of tools and strategies while translating. Some frameworks for learning audiovisual translation promote engagement by emphasizing the benefits of using social connections within the learning environment as reference tools and ways to learn [1]. Non-professional translators typically need support in discerning slang, idiomatic language, and single lexemes. Having access to online dictionaries and materials for checking collocations can greatly assist in the subtitling process. An important aspect of Munro's [12] framework for crowdsourcing translations for relief efforts after the 2010 earthquake in Haiti included an interactive chat room in which volunteer translators discussed lexical items, regional slang, abbreviations, spelling variations, and locations. It also allowed new volunteers to connect with experienced participants and provided a space for people to collaborate.

3. APPROACH

Eight interviews were conducted at the start of this project with professionals in translation, digital materials archiving, and copyright at a major university, including staff members at a language resource center. The expert respondents worked both with professional and non-professional translators on a variety of digital and analog formats.

The language resource center specialized in managing workflows for periodic three-day translate-a-thons in which large groups of volunteers, typically non-experts, translate text and audiovisual materials for local organizations. Interviews with employees from the center provided insights about managing large-scale translation projects, recruiting and motivating translators, and creating workflows for multi-stage processes. Other respondents worked with educational and cultural video materials in libraries or online courses. Several themes emerged in these meetings as well, primarily related to challenges with serial translation, copyright, and hand-offs of a single text by multiple participants.

Following this, we worked with members of the Punjabi community in the region to examine the practicalities of translation work, the availability of translatable resources, and means for initially recruiting volunteers. Our envisioned online system relies on non-professionals working simultaneously on multimedia artifacts. Discussions with the community were used in conjunction with insights from other interviews to explore means of organizing pools of remote translators and identifying events and locations to recruit potential volunteers.

Coupling these interviews with comparative analysis of various crowdsourcing, video repository, and translation systems led us to build over an Amara framework. Amara is a specialized service that offers a robust environment for shared subtitling. It's used by the language resource center, a global open education initiative at our university, and other comparable volunteer-based translation systems. These include Global Voices, a citizen journalism organization that leverages volunteers to translate stories into more than 30 languages, and TED, a nonprofit committed to putting on international conferences that are posted online and translated into over 100 languages by volunteers.

4. DESIGN DECISIONS

The design artifact was built to provide a basic entry point for three types of users – casual observers, translators/reviewers, and administrators. Administrators manage the content and user roles, volunteers contribute translation content and resources, and casual observers are either consumers of video content or potential volunteers. The site aims to encourage people with varying levels of Punjabi language ability to get involved. Strong Punjabi-English bilinguals can actively engage in translating, respond to forum inquiries, and play moderating roles while those with weaker skills or an interest in practicing the language may participate by taking on smaller tasks.

4.1 Framework

We chose to build using the Drupal content management system based on ease of integration with YouTube and Amara. It supports creation of user types with varying permissions and capabilities, complex multi-step workflows involving multiple users, and seamless integration of rich audiovisual content. The main advantage of incorporating Amara is that it breaks down the subtitling process into drafting, syncing/timing, reviewing, and editing by different users. Figure 1 depicts the Amara interface for someone starting a new set of subtitles.

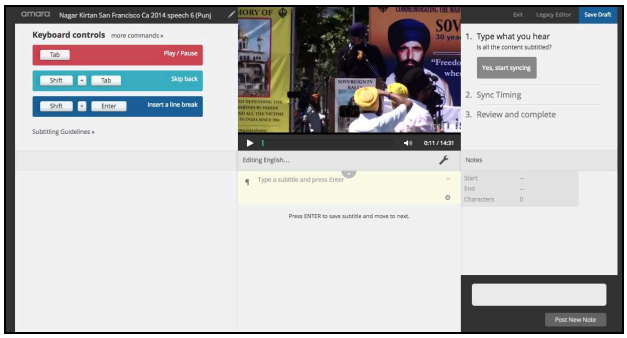

Figure 1. Amara subtitling editor

4.2 Architecture

The site architecture consists of several main pages: video collection browsing, individual video pages with metadata, an explanation of the project with copyright information, instructions and translation resources, an area to submit content, and new translator sign-up. The video collection on the home page displays differently depending on which mode participants choose to operate in—as a general browser, a translator, or a reviewer. This contributes to easier navigation for people interested in doing a specific type of task based on their availability and language skills. Figure 2 shows a snapshot of the home page.

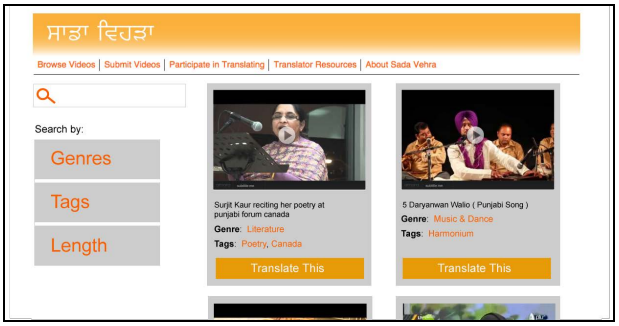

Figure 2. Home page video browsing

The instructions and resources area is made up of translation aides for users. The forum is emphasized to encourage people to first consider reaching out to the network for help. In addition, there are links to Amara-specific documentation, online dictionaries, translation tips, and an instructional video that walks through the process of translating on the site.

Figure 3. Individual video page

Individual video pages serve to display metadata related to the content and encourage users to work on translations and reviews where they're needed. They provide a description with any notes from the translator, show genre and subject tags, and have direct links to the Amara editing interface. Figure 3 shows a sample video page for content that has not yet been translated.

4.3 Content

Members of the Punjabi community in the region have expressed most interest in seeing video content related to Punjabi culture on the site. In the initial phases of the project, this has included but is not limited to music, poetry, folk stories, and food. We hope to gauge availability of and interest in different types of materials, to motivate a wide range of participants to translate videos they have a connection to, and to establish a malleable framework that could be used for other more specific content types in the future.

We intend for the workflow and materials to promote ownership of the system by the crowd. The site design incorporates a mechanism for authenticated users to submit content. One of the biggest challenges many of our interviewees faced with large-scale projects involving multiple people and translation assets was too much reliance on manual data procedures. Offloading some input of new materials and metadata onto users and automating content tracking helps alleviate some of these burdensome and error-prone processes.

The site's collection comprises two main types of videos, all hosted on YouTube. One subset is licensed under Creative Commons, meaning derivative works such as translations can be created without further written permission from the content owners. The other subset contains videos with standard YouTube licenses that require written permission from their owners before they can be published on the site with translations. We decided after interviews with a copyright professional and others working on translating open educational materials that incorporating both types into the system would be ideal.

5. FINDINGS

Some of the challenges we found are not unlike those common for other languages. We break these into three basic categories related to language, infrastructure, and the user interface. First, there are a few issues specific to language structure, variation, and the nuances around the specificities of Punjabi. The quality of the media artifacts also impacts some of the design decisions related to the translation platform. Finally, the choice of interface for the site is intended to facilitate an engaging and productive experience for users who wish to contribute or simply browse.

5.1 Language Variation

Idioms were challenging to translate, and the interpretive meaning could differ from translator to translator. Likewise, the domestication based on regional dialect introduced further challenges since there were distinctions not only in Pakistani and Indian versions, but also in sub-regional dialects. There was some domestication by native English speakers that affects the social power dynamics of language relations, although that seems to have been more common in past translation with trained professionals. This site's aim is not to promote the spread of English as a world language, but rather to put the focus on Punjabi language and culture and use English as a way to spread them to more people [11]. There is a need therefore to build redundancy and triangulation into the system.

Organization names, honorifics and addresses, dates, and religious terminologies were sometimes challenging to translate, as translators had no access to a commonly agreed upon schema for such terms. In addition, kinship terms in Punjabi convey much more specific semantic information, and this specificity is mostly lost when they're translated into English.

Another challenge we found was that users were typically stronger in one or another language, frequently needing to reference language resources. The lack of easy and reliable dictionaries online was sometimes a problem for users as they worked on translation. One crucial word that can't be discerned sometimes

prevents understanding of an utterance that can lead to unclear or lower quality translations.

5.2 Media Infrastructure

Low-quality audio or the presence of excessive background noise during dialogue can pose a great challenge for translators [4]. Misunderstanding single lexemes and longer stretches of connected speech was an issue often related to the nature of the audiovisual materials.

While we were able to conduct AV translation work in the United States where bandwidth was not an issue, we expect that a large amount of the work will need to be done in Punjabi speaking regions where both the bandwidth speed and payment plans where users are charged on megabytes of data downloaded are likely to be a challenge.

5.3 Interface & Workflow

Clear and concise instructions guide users through essential elements of the interface before and during work on an item. Still, managing handoffs between users is likely to be a major challenge, especially when bringing new translators into the fold. There is a wide variance in time periods for translating a specific asset both based on what is in a media resource and on the fluency of the translator.

Giving the crowd ownership of the platform and their own work is central to creating a sustainable environment in which people are willing to contribute regularly. A forum for people to exchange ideas can be critical, even with professional quality translation services. We find that users express a need to ask about dialectal variation and idiomatic or slang language that may not be used everywhere. During the language resource center's translate-a-thons, volunteers often express the value of being able to communicate with friends and family members elsewhere who are native speakers that can provide important insights about certain words, phrases, or the style of speech.

6. FUTURE WORK

The next steps in this project involve conducting user experience testing and iterating on the existing prototype to fully develop all pages and the workflow of the system. A large part of this process is to make the reviewing and editing of draft translations more intuitive. The browsing experience will be refined based on what seem to be the most salient ways to organize and categorize the videos. The admin structure will also be improved to facilitate efficient management of participants and resources.

The project involves in-depth interviews with translators to understand motivations and the translation experience. Quantitative data includes tracking demographics alongside time spent per item, repeat usage, and handoff efficiency. We will also track participation and use of forums.

Finally, our goal is to facilitate larger corpora of data to support equivalency theories for multilingual speakers' translations. Crowdsourcing is an important means of generating datasets of translations to advance understandings of equivalencies between language pairs [17], and currently there exists little literature focusing on equivalency between English and Punjabi. Our goal for Sada Vehra is to provide a prototype to support the creation of new online forums for translating languages with limited online content, but a fairly large population of online speakers.

7. REFERENCES

[1] Amador, M., Dorado, C., and Orero, P. 2004. e-AVT: A perfect match: Strategies, functions and interactions in an online environment for learning audiovisual translation. In *Topics in audiovisual translation*, P. Orero, Ed. John Benjamins Publishing Co., Philadelphia, PA. 141-153.

[2] Ambati, V., Vogel, S., and Carbonell, J. 2012. Collaborative workflow for crowdsourcing translation. In *Proceeds of the ACM CSCW 2012 Conference on Computer-Supported Cooperative Work* (Seattle, WA, February 11-15, 2012). ACM, New York, NY. 1191-1194

[3] Brabham, D. C. 2013. Crowdsourcing. The MIT Press, Cambridge, MA.

[4] Bugocki, Ł. 2009. Amateur Subtitling on the Internet. In *Audiovisual translation: Language transfer on screen*, J. Díaz Cintas & G. Anderman, Eds. Palgrave Macmillan, New York. 49-57.

[5] Danet, Brenda, and Susan C. Herring, eds. The multilingual Internet: Language, culture, and communication online. Oxford University Press, 2007.

[6] Díaz-Cintas, J. and Anderman, G. 2009. *Audiovisual translation: Language transfer on screen.* Palgrave Macmillan, New York.

[7] Ethnologue. 2014. Languages of the World: Panjabi, Eastern. Retrieved December 20, 2014 from http://www.ethnologue.com/language/pan

[8] Ethnologue. 2014. Languages of the World: Panjabi, Western. Retrieved December 20, 2014 from http://www.ethnologue.com/language/pnb

[9] Gottlieb, H. 2004. Language-political implications of subtitling. In *Topics in audiovisual translation*, P. Orero, Ed. John Benjamins Publishing Co., Philadelphia, PA. 83-100.

[10] Martínez, X. 2004. Film dubbing: Its process and translation. In *Topics in audiovisual translation*, P. Orero, Ed. John Benjamins Publishing Co., Philadelphia, PA. 3-7.

[11] Munday, J. 2012. *Introducing Translation Studies*. Routledge, London and New York.

[12] Munro, R. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation* (Denver, CO, October 31, 2010). AMTA.

[13] Neves, J. 2004. Language awareness through training in subtitling. In *Topics in audiovisual translation*, P. Orero, Ed. John Benjamins Publishing Co., Philadelphia, PA. 127-139.

[14] Pal, J. et al (2012). Local-language digital information in India: challenges and opportunities for screen readers. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. (Atlanta, GA, March 12-15, 2012). ACM. 318-325

[15] Pérez-González, L. 2014. *Audiovisual Translation: Theories, Methods and Issues.* Routledge, Taylor & Francis Group, London.

[16] Sánchez, D. 2004. Subtitling methods and team-translation. In *Topics in audiovisual translation*, P. Orero, Ed. John Benjamins Publishing Co., Amsterdam/Philadelphia, PA. 9-17.

[17] Zaidan, O. F. and Callison-Burch, C. 2011. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Portland, OR, June 19-24, 2011). ACL. 1220–1229.