Connecting Documentation and Revitalization: A New Approach to Language Apps

Alexa N. Little
7000 Languages
12 Murphy Drive
Nashua, NH 03062
alittle@7000languages.org

Abstract

This paper introduces 7000 Languages, a nonprofit effort to adapt commercial language-learning software for free use by endangered language groups. We discuss the advantages and challenges of our approach and suggest ways of converting corpus and archive data into language-learning courses. We also demonstrate the capabilities of the software, both for the revitalization and the documentation of endangered languages. Finally, we discuss ideas for future expansion of our programs and seek feedback from others involved in endangered language work.

1 Introduction

Many endangered language communities are interested in producing "language apps": digital tools for teaching and learning their languages. Although these language apps are useful for language revitalization efforts, building them requires considerable time, funding, and technical skill. Even successful community projects may lack the resources for future updates to their language apps.

Researchers who study endangered languages also want to provide digital tools, so that their research and data will be helpful to communities. Creating and maintaining an entire language-learning system, however, is more than a researcher can reasonably accomplish.

The current compromise is the digital archive. In a digital archive, language data is secure, accessible, and generally portable. XML-formatted data could even, with some manipulation, be integrated into a language app. However, the time and expense of building the rest of such an application remains a challenge for both sides.

We introduce a nonprofit effort, 7000 Languages, that seeks to resolve this problem. Our approach uses technology donated by the language-learning industry to produce free commercial-grade language apps in partnership with endangered language advocates.

2 The Programs

We have organized our approach into two programs – one from a revitalization-first perspective, and one from a documentation-first perspective. Below is a brief description of these two programs, followed by an in-depth discussion of how projects starting from either perspective will ultimately benefit both sides.

2.1 Partnership Program

Our Partnership Program is a free program intended for groups, such as endangered language communities, who have a revitalization-first perspective. The system works as follows:

7000 Languages has an agreement with a for-profit language-learning company, Transparent Language. This agreement allows 7000 Languages to use Transparent Language's internal tools to develop online courses for endangered and low-resource languages. The tools include:

- a program that converts an Excel template filled with language content (text, images, and recordings) into a functioning course
- a program for designing a custom unit, lesson, and activity layout
- a graphical user interface (GUI) for creating individual lessons, ideal for those with limited technical experience (see Figure 1)

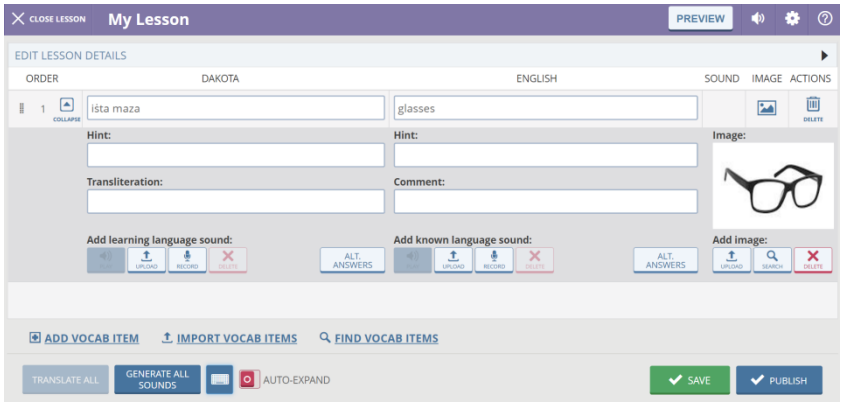

Figure 1. A GUI for lesson authoring.

7000 Languages trains interested groups, called Partners, to use these tools. The Partner decides what they want to teach, and they retain ownership rights of all the content they create. Once the course is finished, 7000 Languages contacts Transparent Language, who sets up a "node" for the Partner in their systems. This publishes the course online and lets the Partner create user accounts. The Partner can add as many users as they want, for free. The course will also be made available, for free, on the 7000 Languages website and published through Transparent Language's library and education subscriptions.

Published courses present an opportunity to produce documentation of the language. Because the courses accept and store language data from the tools in XML format, data can also be exported from the course in XML format. The XML entry for each item contains the phrase in the target language, a translation in the source language, a reference to the target language audio file, and a reference to the source language audio file (if available). See Figure 2 for a partial sample of an XML file exported from a lesson.

Once exported, the XML files can be submitted for archival, used to form a corpus, or reformatted and used in another digital application, such as a searchable dictionary. It is our hope that this will allow endangered language groups to focus on revitalization without having to sacrifice documentation and other uses of their language data.

```
- <card>
    <side1_phrase>blue</side1_phrase>
    <side2_phrase>pelung</side2_phrase>
    <guid>{26B1AD8D-7D64-49C4-BEB0-ADABB9F81418}</guid>
    <list_position>3</list_position>
    <side2_sound url="sounds/learn002.mp3"/>
    <is_video_sound_used>false</is_video_sound_used>
    <is_video_auto_played>false</is_video_auto_played>
  </card>
- <card>
    <side1_phrase>brown</side1_phrase>
    <side2_phrase>coklat</side2_phrase>
    <guid>{DB6B1A48-050F-4851-94EB-79A68C1968D3}</guid>
    <list_position>4</list_position>
    <side2_sound url="sounds/learn003.mp3"/>
    <is_video_sound_used>false</is_video_sound_used>
    <is_video_auto_played>false</is_video_auto_played>
  </card>
- <card>
    <side1_phrase>gray</side1_phrase>
    <side2_phrase>klau</side2_phrase>
    <guid>{C12DB5E5-B9EE-4943-9FD9-2A47F6BFCCDC}</guid>
    <list_position>5</list_position>
    <side2_sound url="sounds/learn004.mp3"/>
    <is_video_sound_used>false</is_video_sound_used>
    <is_video_auto_played>false</is_video_auto_played>
  </card>
```

Figure 2. Vocabulary items from a Balinese lesson, exported as an XML file.

2.2 Archives Program

Our Archives Program is a free program intended for groups, such as field linguists, libraries, and archivists, who have a documentation-first perspective, such as field linguists, libraries, and archivists. It assumes that some language materials already exist, and that both the group holding the data and the larger language community are interested in a language app. The focus of the program is manipulating the existing data to be usable in a language app.

Our goal for the Archives Program is to give documentation-focused researchers the opportunity to contribute, without much additional work, to language revitalization efforts in the relevant community.

The Transparent Language courseware (i.e., the technical components of the language app that make it interactive) generally requires content in a specific XML format. If the linguist or archivist who holds the data is familiar with rewriting XML, we ask them to rewrite their data

to match those specifications. If not, 7000 Languages can often create the appropriate conversion programs. (In the future, we hope to create standard scripts capable of translating XML and other text schema commonly used for language documentation into courseware-compatible versions.)

Once the data has been reformatted into a compatible XML file, and we have verified it will integrate properly with the courseware, we can create interactive lessons simply by importing the data. Just as in the Partnership Program, those lessons can then be published online, and the community given control over a "node" where they can create unlimited free user accounts. The course will also be distributed by 7000 Languages, if the community permits, and made available through Transparent Language's library and education subscriptions.

3 The Software

We will not attempt to show the entire functionality of the software in this short paper. Rather, we present examples that demonstrate the broad capabilities of the technology.

7000 Languages Partners can choose from over 40 different activity types to create their courses. These include matching, multiple choice, and variations on reading, writing, listening, and speaking practice. Figure 3 shows examples of several core activities. The technology also supports simulated conversations (in which the learner provides audio for one of the roles), videos, and text-based reference materials. Lessons usually include assessments, and quizzes are also available as a type of practice drill.

The system allows learners to track their progress. Each vocabulary item that a learner studies is stored in her "learned items." Over time, learned items fade and must be "refreshed" with practice activities. The learner can also consult her list of learned items to see what words she has learned. This list is searchable both by source and target language.

Teachers can also use the system to monitor their students' progress; from the Instructor Portal, they can assign specific lessons and track student activity.

Finally, the online course connects to a smartphone app, which contains flashcard-based activities with the content from each lesson. See Figure 4 for an example. If the user's phone is online, the learned items automatically synchro-

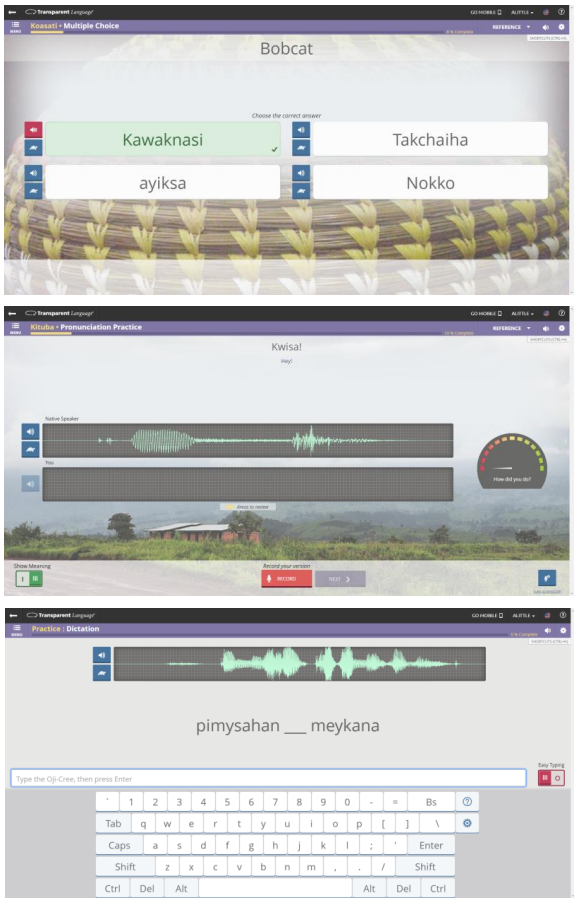

Figure 3. Sample activities on the web app.

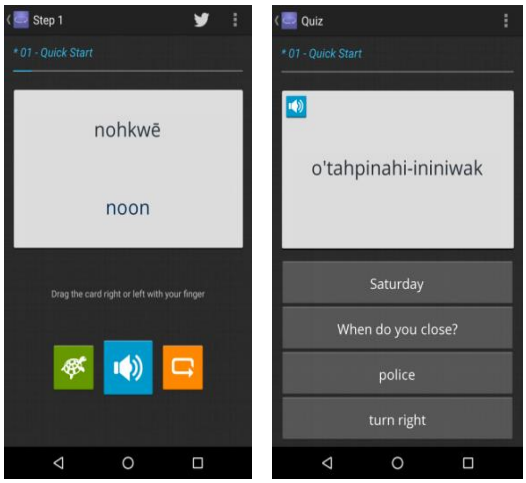

Figure 4. Flashcard-style activities on the mobile app.

nize between the online and mobile versions of the course. The flashcards can also be downloaded, then used even without an Internet connection. The next time the user is online, she can press the "sync" button to have her learned items synchronize between devices.

4 Benefits

We see several ways that our approach could benefit members and supporters of endangered language communities.

4.1 Functionality

As shown in Section 3, the software donated by Transparent Language offers a wide range of activity types and contains options for progress reporting, assessments, and mobile learning. All this functionality was originally funded by commercial interests and intended for commonly-taught languages. However, our approach allows endangered language communities to take advantage of this rich functionality as well.

4.2 Stability and Portability

As endangered language advocates strive to produce digital documentation and language apps, deprecation and portability has increasingly become an issue (Bird and Simons, 2003). An app that works on today's computer or smartphone may not work on the next generation model. And, with the speed of operating system updates, language apps must be updated frequently to remain functional. Finally, consumers can choose between several different operating systems, different types of smartphones, and even different web browsers. Ensuring that a language app will function on all of these systems is a significant challenge.

Because the courses produced under 7000 Languages rely on the same technology as Transparent Language's commercial courses, they benefit from the same updates. This reduces the financial and technical burden on communities to keep their individual language apps up-to-date, and helps ensure that the apps will remain usable across most common platforms.

4.3 Reduced Cost to Community

If the technology involved in creating a language app is available at no cost, communities can dedicate whatever resources they have to other purposes, such as hiring teachers, paying linguistic consultants, or developing additional learning materials.

4.4 Documentation and Revitalization

Because our approach allows language-learning lessons to be downloaded as archivable XML, and materials archived as XML can likewise be converted into language-learning lessons, it offers the potential for increased collaboration between revitalization-focused and documentation-focused groups. For example, an endangered language community could create a language course, then export the material for archival and linguistic analysis. Meanwhile, a linguist could conduct fieldwork, save their language documentation in XML format, and have that XML data converted into a usable language app for the community.

5 Challenges

Our work has been impacted by several challenging factors, and we are interested to hear the perspective of other endangered language advocates on these issues.

5.1 Licensing and Control

In order to use the technology of a for-profit language-learning company, we agreed to certain conditions. They are as follows:

- Transparent Language technology may only be used by 7000 Languages to produce courses for low-resource languages (i.e. not for commonly-taught world languages).
- Finished courses will be published on Transparent Language's library and education services. This means that users who pay to subscribe to Transparent Language courses will receive access to these courses, also.
- If a 7000 Languages Partner chooses to sell the course they created, they must pay a 20% royalty to Transparent Language for the use of its technology. However, if they distribute the course for free, there is never any royalty required.
- If Transparent Language ever chooses to sell a course created by a 7000 Languages Partner as a standalone course, they will pay a 20% royalty to the Partner.

Many communities are willing to accept these conditions in order to gain free use of Transparent Language's technology and the benefits we described in Section 4. However, we understand that some communities are not comfortable with a for-profit company controlling any aspect of their language, or with their language being widely available for anyone to learn. In such situations, our approach is not a good fit.

5.2 Sensitive Material

We have been approached by communities who wish to restrict access to some of their content, for cultural reasons. While we understand the need to both preserve this material and teach it only to the appropriate people, the technology does not have the sophisticated access controls required to make that possible. At the moment, we encourage these groups to design lessons with material that *can* be shown to the general public, and to use other resources to teach and archive sensitive material.

5.3 Grants

Although our approach greatly reduces the cost to endangered language groups by providing free access to a technological framework, it does not completely eliminate the costs of producing a language app. Designing a curriculum, making text and recordings, and learning to use the tools requires time and effort. Some of the groups who work with us have grant funding, but many do not—which means that they take on this work as volunteers.

5.4 Internet Access

Many of the communities facing language loss may also lack access to consistent, high-speed Internet service. Poor Internet infrastructure remains a problem in remote areas of the United States and Canada, even after 20 years (Carpenter et al., 2016). The inability to reliably access the Internet is a barrier for communities who are otherwise enthusiastic about adopting new technologies (Carpenter et al., 2016).

Transparent Language technology does function offline in the form of a mobile flashcard app. However, the fully-featured software generally requires a computer and an Internet connection, or an iPad onto which the software can be downloaded.

Because a computer and an Internet connection are used by Partners to develop courses in the first place, this has not yet been a major concern. However, some Partners may be unable to distribute their courses as widely as they wish because community members lack reliable Internet access. Furthermore, as the Archives Program expands, the course creators and course users are increasingly likely to be different groups altogether, with different levels of access to technology. We have anticipated these problems and are considering possible solutions now, before they become an active concern.

6 Future Work

The mission of 7000 Languages is to connect endangered language communities with the technology they need to teach, learn, and revive their languages. We recognize that the difficulty of creating language apps is not the only technological barrier that these communities face. In this paper, we have mentioned some possible future directions for increasing the impact of our program. We intend to make data exported from completed courses compatible with other technologies, such as dictionary apps or Natural Language Processing programs. We also plan to develop scripts to smooth the conversion process between documentation and language app. In planning our next steps, we look to the language revitalization and documentation field for suggestions, constructive criticism, and opportunities for collaboration.

7 Conclusion

In this paper, we introduced 7000 Languages, a nonprofit effort that uses technology donated by the language-learning industry to create free endangered language apps. We discussed the current approach of 7000 Languages, we described some features of the software, and we sought feedback from other endangered language advocates.

Acknowledgments

The programs described in this paper were made possible by the board of 7000 Languages, Transparent Language, and the individual donors who support our organization. Thanks to R. Regan and C. Graham for their assistance.

References

S. Bird and G. Simons. 2003. *Seven dimensions of portability for language documentation and description.* In Bojan Petek (ed.), Portability issues in human language technologies: LREC 2002.

Jennifer Carpenter, Annie Guerin, Michelle Kaczmarek, Gerry Lawson, Kim Lawson, Lisa P. Nathan, and Mark Turin. 2016. *Digital Access for Language and Culture in First Nations Communities*. Knowledge Synthesis Report for Social Sciences and Humanities Research Council of Canada. Vancouver, October 2016.