Language revitalization in the digital age: building bridges, empowering communities

Kevin Scannell, Saint Louis University

This is the text of a paper I presented at the Walter J. Ong Symposium on Digital Humanities at Saint Louis University, March 23, 2016.

Slide 1: UNESCO map

We're living in an era of unprecedented language loss. Approximately 7000 languages are spoken today, a more or less stable number over the last 1000+ years. Estimates vary, but most experts expect at least half of these will no longer be spoken by 2100. The image here is from the UNESCO Atlas of the World's Languages in Danger, and it shows that language loss is a worldwide phenomenon. No two contexts are alike, but for many of these languages, intergenerational transmission has stopped, and the language has given way to English, French or another global language in certain domains: the home, in schools, or in obtaining government services.

I come to this work as a speaker of Irish (aka Irish Gaelic), which has maybe 25,000 native speakers remaining and is listed as "definitely endangered" in the UNESCO atlas.

Slide 2: Digital language revitalization

Communities around the world are working to revitalize their languages, which means using them again in domains where they've retreated, and adapting them to new domains. One very important domain, especially to me as a computer scientist, is **technology, computing, and the internet.** Indeed, the internet presents **fantastic opportunities** for language revitalization, but **also some obstacles.** I'd like to discuss both.

A common issue in minority language communities is that they don't form a critical mass in any one place; native speakers may move away from the geographic "home base" of the language. The Irish diaspora is a good example of this. The internet provides a vehicle for us to overcome this obstacle, by reconnecting diaspora speakers with native speakers. In Ireland this can include speakers living outside the Gaeltacht in cities like Dublin or Cork, but also those of us who speak the language and live overseas. There are similar dynamics among some native communities in North America, many young people moving off of the reservation to urban areas; and also for many African languages with a growing number of diaspora speakers using their languages from Europe, N. America, etc.

The image you see here is one I created about three years ago; it shows **Irish-language "conversations" on Twitter** in cases where I was able to geolocate both the sender and receiver. It shows a global community connecting with native speakers in the Gaeltacht.

I just finished a three-year NSF project which involved a **wide “crawl” of the web** to collect material written in as many languages as possible. The real aim of the project was to create databases to facilitate linguistic research and the development of natural language processing software, but an interesting side benefit was the fact that it provides **the best sociolinguistic survey of the web to date**.

Between 2500 and 3000 languages have some online written presence. Most of the others have never been written at all! And even more surprising, perhaps, is that majority of the 3000 that are online consist entirely of Bible translations or evangelical material of one kind or another. My best estimate is that **less than 1000 languages have a writing system in regular use by the language community itself**. Written language is truly the exception.

Slide #3 Indigenous Tweets

I’m particularly interested in social media sites like Facebook and Twitter as vehicles for language revitalization. Tweets have been sent in somewhere between 400 and 500 languages; in 2011 I created a web site called “Indigenous Tweets”, pictured here, that tracks everyone using Twitter in an indigenous or minority language. I started with 35 languages in 2011, and we’re now tracking 180 and have found more than 34 million tweets written in these languages. In many cases it paints a picture of truly resurgent language communities, not “dead” or “dying” languages. The site shows statistics on how many tweets each user has sent, the percentage in the language, number of followers, etc., in order to serve as a kind of “menu” for who to follow in your language. Another nice feature is that I compute “trending topics” on a per language basis; it’s extremely unusual for trends in a minority language to bubble up and appear as trends on Twitter.com.

Slide #4 Yuchi

Here’s a tweet from the Euchee language project. Yuchi is spoken in Oklahoma. Only five fluent speakers remaining, all elders, but there are efforts underway to pass the language on to the next generation, and there are some young people actively tweeting in the language. It’s an interesting case also because it’s the only language in the world that uses the @ symbol as part of its alphabet, which causes some problems in tweets as you see here!

Slide #5 Obstacles

It’s not all good news. There are some huge obstacles in terms of getting people to use their languages online. Issues of poverty, illiteracy, lack of connectivity, historical trauma, and so on are global challenges and exceedingly difficult to overcome. In cases where communities are well-connected and eager to use their languages online, there’s still technical work and outreach to do. Most of my activities involve these issues:

- Social media primarily use written language; can we create tools for the 80% or more of languages that are not written?

- “Discoverability” in a vast sea of English online (Indigenous Tweets was created in part to overcome this obstacle)
- Normalization: “the computer is in English”, which is a kind of chicken-egg problem (we’ll discuss software localization in a moment)
- Incredibly fast moving landscape
- Language-specific technical obstacles: fonts, input methods
- Terminology creation
- Lack of technical skills in language communities
- Corporate control of platforms

Slide #6 Software Localization

One place where we’ve been able to have a real impact is by **translating software interfaces**; this means standalone software packages, games, web browsers, office software, and a number of widely used websites. You see screenshots here of the **Twitter** interface translated into Irish, which I completed together with three other speakers last year, and the **Firefox** web browser, which we’ve had in Irish since 2004. This kind of immersive, monolingual environment is absolutely essential for those communities trying to establish immersion education programs in their languages; this simply isn’t possible if books, learning materials, and computer programs are always in English!

Slide #7 Input methods
These are basic tools that majority language speakers take for granted. Predictive text, as shown here on the left for Manx Gaelic. Or spelling and grammar checkers for your languages; the screenshot on the right is my Irish language spell checker running in Firefox, in a fully-translated version of Gmail (I project we launched two years ago in Dublin at Google’s Irish headquarters).

Slide #8 Terminology
Something else that majority language speakers take for granted, since terms tend to emerge organically from the crowds of people using them everyday. I’m a member of the subcommittee of the Irish language terminology committee charged with creating terms for computing and technology. It’s a tremendous challenge, but also good fun. Here are a couple of examples that draw on native elements:

- “Turscar” (cast-up seaweed = spam)
- “Ubh chuaiche” (cuckoo’s egg = cookie)

Every language must create new terms if speakers want to engage with modern concepts and technology in their language. In practice it’s a huge problem for communities just starting on software localization for example. One thing I’ve tried to encourage is the borrowing of metaphors among indigenous language groups. Sometimes the metaphor in English simply isn’t suitable for the way a particular language community conceptualizes something; it can be helpful to hear how other groups have created terms, since sometimes a particular metaphor will “click” and suggest the perfect term in another language (especially related languages, but even if not). Some examples I like, created by Edmond Kachale for Chichewa:

- Ulalo (bridge = link on a web page)
- Mkute wa tsambe (page leftovers = cached page)

Slide #9 “Never Beg to save the language”. This quote is from a piece by the late Darrell Kipp, a hero of the Blackfeet language revitalization movement.

This is maybe the most significant obstacle we face in trying to make endangered languages visible in computing and technology: in most cases, this requires the cooperation of big tech companies. Earlier I mentioned successful collaborations with Twitter and Google in Dublin for the Irish language, but we’re a special case in many ways. Looking more broadly, it’s clear that the big tech companies don’t really care about endangered languages, despite what they might say publicly.

Facebook in particular has stated that they’d like to make their site available in “all the world’s languages”. And to their credit, they created a nice system that allows users of the site to translate into more than 100 languages (though on a volunteer basis!) But they’ve only added a handful of languages in the last 5 years, despite huge demand, petitions, etc., and crowds of people willing to do the work for free!

In 2010, a friend named Neskie Manuel created a piece of software that runs in your web browser and acts as a kind of “Facebook overlay”; allowing one to provide a translation of the site without needing Facebook’s permission. I’ve worked with more than 40 language groups to use Neskie’s overlay to provide partial translations of Facebook. What you see here is NOT an official translation; instead, Facebook is serving content in English and the overlay translates the navigation, “Like”, “Share”, etc.... in this case, into Chichewa, thanks to a translation by Edmond Kachale.

Slide #10: Themes

- Open source software and open data
- Community v. corporate ownership of resources
- Language-independent tools when possible
- Volunteer contributions from the community
- Capacity-building; code-ins, hackathons

Slide #11: Praistriúchán/Scymraeg

What you see here are some famous examples of “machine translation fails”. I could provide you with hundreds of others. The top-left is obvious enough. The bottom left is similar: the Welsh on the sign means “I am not in the office at the moment. Send any work to be translated.”! On the right is a notice from the Dublin airport. “Please be patient...” In English gets translated into Irish using the word for a medical patient “othar”!

Praistriúchán and Scymraeg are both portmanteau words. The first is Irish, a mix of “aistriúchán” (translation) and “praiseach” (a mess). The second is Welsh, a mix of “Cymraeg” (Welsh) and the English word “scum”. It’s telling that there are words for this concept in Irish and Welsh and no similar term (and no real need for one) in English! Indeed, there’s a Scymraeg Flickr group with more than 500 similar images uploaded.

Machine Translation researchers and Google in particular have the idea that they’re performing a great service on behalf of endangered language groups. But here’s the thing -- Google can, in principle, add a language to Google Translate without hiring, consulting with, or even speaking to a single speaker. It’s simply a matter of collecting enough parallel text and then creating appropriate statistical models. This is problematic to say the least.

Slide #12: Maybe we should listen to the communities themselves.

On this slide I’ve linked a number of articles and blog posts that reflect the views of some indigenous languages speakers on machine translation and Google Translate in particular.

Community Response: Hawaii
“Before, it was illegal to speak Hawaiian and now it’s at everyone’s disposal. I feel like making it so easily accessible makes it more vulnerable for exploitation... [one] should actually take the time to learn it from someone who has spoken it or knows it and not an automated internet source that does it all for you”

“Because United States assimilationist policy attempted to eradicate Hawaiian language and culture, our language had and continues to have a need for revitalization – more speakers... [but] there are inherent problems in learning a language outside of its cultural context, and the format of Google Translate ensures that this problem can never be fixed”

They go on to talk about the famously untranslatable “aloha”

Scotland: “Whatever the intentions of the developers, people will mis-use such a system. I have put together a few annotated photos which illustrate the scale of the disaster in Ireland... From school reports to official government websites, there are few places where students, individuals or officials trying to cut corners have not used Irish translations of Google Translate in ways they were not intended to be used.”

Ireland: “Do Minority Languages Need Machine Translation?” In this third piece, the author argues “no”, and specifically that money spent on machine translation would be much better spent on tools like the ones I mentioned above, that help speakers of the language use the language productively: spelling and grammar checkers, software localizations, monolingual and bilingual dictionaries

These discussions raise broader questions: untranslatability, connection of language with traditional culture, and “ownership” of language.

The last paper listed, by linguist Jane Hill, while not directly addressing translation, is relevant to this discussion, and a paper that’s had a great influence on me personally. It is a criticism of the standard discourse one hears from western, usually academic, linguists who advocate for linguistic diversity or preservation of endangered languages:

- Universal ownership (“loss to us all”, “loss to humanity”, “*our* linguistic diversity”, “*the world’s* languages”)
- Hyperbolic valorization (“cultural treasures”, “immeasurable wealth”)
- Enumeration, counting languages and counting speakers (“the very notion that languages can be counted and named may be part of the disease that as affected the linguistic ecology of the Pacific”)

Slide #13: Intergaelic

This is a site I created that’s an attempt to be a “better” Google Translate. It translates only between the three Gaelic languages; Irish, Manx and Scottish Gaelic. These are linguistically quite close, so it’s possible to achieve very high accuracy. But they’re distinct enough that it’s a barrier to communication between Gaelic speakers in Scotland and Ireland.

One focus of the project has been to deliver machine translation results in innovative ways. We do this, for example, by providing live translations of social media content. And on the Intergaelic site itself (as in the screenshot here) you can see the output of a translation from, say, Scottish Gaelic to Irish; unlike Google translate, we just output the original text you put in, and provide the translation as (small!) annotations when they’re needed.

Slide #14

Let me close with a quote that captures my feelings on these topics pretty well! It’s from my favorite Irish language writer, who uses the pen name “Biddy Jenkinson”. She doesn’t allow her books to be translated into English.

“I prefer not to be translated into English in Ireland. It is a small rude gesture to those who think that everything can be harvested and stored without loss in an English-speaking Ireland. If I were a corncrake I would feel no obligation to have my skin cured, my tarsi injected with formalin so that I could fill a museum shelf in a world that saw no need for my kind”