Translating Facebook into Endangered Languages

Kevin Scannell
Saint Louis University, St. Louis, Missouri, USA
[kscanne@gmail.com]

Abstract: Facebook is an incredibly popular social networking site, with more than 900 million users as of March 2012. Many indigenous and minority language groups are turning to Facebook as a way for small and scattered speaker populations to connect with each other online. There are at least partial translations of the site into about 100 languages, including several endangered languages, such as Irish, Northern Sámi, and Cherokee. Unfortunately, Facebook have not added any new languages to their official "Translation App" in more than a year, and there are no signs that they will any time soon. I will discuss a technical solution to this problem, originally due to Neskie Manuel, that allows the site to be translated into any language, in the user's browser, without the need for Facebook's approval or cooperation. We have used this approach to provide new Facebook translations into 29 new languages in cooperation with native speakers.

Social media and endangered languages

Social media web sites like Facebook, Twitter, and Google+ allow users to connect with one another, share content, and communicate online. Hundreds of language groups around the world have recognized the potential these sites have for their language revitalization efforts, both in terms of encouraging language use and sharing techniques across communities.

Consider, for example, the Irish language, which is spoken as the everyday community language only in small areas of Ireland, collectively known as the *Gaeltacht*. Outside of the *Gaeltacht*, there are few places with the critical mass of speakers needed to sustain a speech community. Nevertheless, interest in the language is surging among people living in urban areas such as Dublin and Belfast, and has grown also among diaspora Irish in the UK, Australia, and North America. Through social media, speakers from these scattered communities are able to communicate with each other and with *Gaeltacht* speakers in a way that was impossible even ten years ago. And though we should be careful not to overgeneralize, there are clearly similar dynamics at play with some other European minority languages (Welsh and Basque for example), and I suspect in other parts of the world as well.

As of March 2012, Facebook reported about 900 million active users, and Twitter about 140 million. Both sites have substantial numbers of endangered language speakers. Indeed, we can back this claim up with detailed numbers in the case of Twitter. Last year I created a website called Indigenous Tweets (Scannell, 2011) that tracks everyone using Twitter in an indigenous language (many, but not all, of these are endangered languages). The idea was to allow speakers from small language communities to discover each other more easily on a site where endangered language voices are easily drowned out by the hundreds of millions of tweets each day in English, Japanese, Spanish, etc. To date we have identified more than 7.5 million tweets by 46000 users written in 136 indigenous languages.

The social dynamics on Facebook and Twitter are somewhat different. On Twitter, connections can be unidirectional; that is, users can follow their friends, but also strangers such as politicians or celebrities, who are unlikely to follow back. Many people find Twitter be a good way to "meet" new people with similar interests. The 140 character limit and the informal register make Twitter especially suitable for language learners and semi-speakers who are able to use the bits of language they know while learning from more fluent speakers.

On the other hand, Facebook connections are typically bidirectional, and as a consequence people are less likely to connect with people they do not already know in "real life". Endangered language activity on Facebook is often centered around "groups" devoted to a given language, where discussions in or about the language can take place. I belong to groups devoted to Irish, Manx Gaelic, UmoNhoN/Omaha, Nawat/Pipil, and I have seen dozens of others with similar aims.

Facebook has some advantages over Twitter for language revitalization activities, for example, better multimedia support and clearer threading of conversations. Also, if you post messages in more than one language, it is possible to maintain separate lists of friends and target posts in a given language to only those friends who can read that

language. On Twitter, all of your followers get all of your tweets, and so some multilingual tweeters have resorted to creating multiple accounts, one for each language they speak.

As a final remark, it is worth remembering there is no "silver bullet" in language revitalization. Social media will not "save your language", despite claims to the contrary in the mainstream media, but we believe they can play an important role in broader revitalization efforts by breaking down the idea that indigenous languages have no place in the world of computing and technology, and by getting people to use their language more frequently.

In the remainder of this paper, we will describe our ongoing effort to provide translations of the Facebook interface into as many endangered languages as possible.

Official Facebook Translations

Given suitable fonts and keyboards, it is possible to create *content* in virtually any language on any of the sites mentioned above. That said, the menus, navigation, and prompts on the sites themselves are only very rarely made available in endangered languages. So while you may be able to post status updates to Facebook in Ojibwe or Kashubian, you are at present forced to use the site in English or some other major language.

To their credit, Facebook have created an innovative and powerful system for translating their site, relying almost entirely on volunteer translators. One nice feature of this system is that translations can be done "inline". This means that if, during your normal day-to-day use of Facebook, you see an untranslated message, you can click it and provide a translation on the spot. This approach provides a measure of instant gratification that is unavailable in most traditional software translation contexts, where translations must be submitted to developers, who compile test versions, which are then made available to the translators for proofreading and testing. Inline translation also helps with quality since the messages to be translated appear "in context" on the site.

Quality is controlled through a voting system; in addition to supplying new translations, one is able to vote other volunteers' translations up or down. A translator's influence is measured in terms of both quantity of translations and their quality, as measured by the voting system.

It is hard to estimate the exact number of words required for a complete translation of the site since Facebook does not expose these global statistics to volunteers, and because the site is changing and growing constantly. By any measure, though, it is a massive undertaking for a small language community. The top 100 most active Irish language translators have, to date, contributed almost 61000 translations (with multiple translations of the same messages permitted), yet Facebook reports that our translation is still only 80% complete.

Spanish was the first official translation released, in January 2008. Since then, at least partial translations have been done for about 100 languages through the translation platform, including a number of languages listed in the UNESCO Atlas (Moseley, 2010) as "vulnerable" (Aymara, Basque, Faroese, Frisian, Welsh) and "definitely endangered" (Cherokee, Irish, Northern Sámi, Rumantsch).

Which brings us to the problem. Despite their stated goal "to make Facebook available in every language across the world" (Haddad, 2009), and despite numerous requests and petitions, no new languages have been added to the official translation platform in more than a year, effectively locking out speakers of the 6800+ unsupported languages.

Unofficial Facebook Translations

In 2010, faced with this difficulty, my friend Neskie Manuel came up with a technical solution that allowed him to do a partial translation of Facebook into his language of Secwepemctsín (Manuel, 2010) using a technology called "Greasemonkey" (Greasespot, 2012). The idea is that a user can install a script in his or her web browser that is activated upon visiting facebook.com, and which acts as an "overlay" in the language of choice.

At one level, this can be viewed as a simple technical trick or a "hack". But at another level I believe this can be a game-changing idea for indigenous language groups wanting to use their languages online. As mentioned above, we

can petition and plead with big tech companies to support our languages, and sometimes they do. Google's search interface is available in about 150 languages, including several endangered languages. But what about GMail, or Google+, or whatever Google's next big thing might be? With each new site or service, endangered language groups must go hat-in-hand asking for their languages to be included. And the long-term prospects for this approach look grim for languages that are of little commercial interest, that lack political clout, and that do not have "insiders" in the tech community. As mentioned, Facebook has stopped taking on new languages. Google has effectively shuttered their "Google in Your Language" program. Even the non-profit Mozilla Foundation, which has a long tradition of supporting minority language translation of its products, has had difficulty scaling up their translation infrastructure to support more than 100 languages.

Neskie's idea means that endangered language groups can take control of their online presence, without being beholden to corporations that do not, when push comes to shove, really care about endangered languages. When Neskie passed away tragically in May of 2011, I decided to continue work on the Facebook project, as a way of honoring his memory and also to keep this important idea alive. In January 2012, I released a first version which updated Neskie's code to work with the new Facebook design, and generalized it to work, in theory, with *any* language.

As noted above, the full Facebook interface is huge, and probably overwhelmingly large for a small language group. One of my main goals was to make the translation feasible for languages that have never attempted software translation before, and for this reason I decided to limit the translations to 200 short messages. As a result, the translation will be far from complete, but by choosing these high-priority messages carefully, we have been able to get all of the important navigation elements and the most common messages seen in a user's News Feed. A translation of all 200 messages provides a convincingly immersive monolingual experience in the endangered language, at least for the most often used pages on Facebook (user profile pages and the Home page which contains the News Feed).

Because the translation will never be 100% complete, I added another feature that allows each language group to select a "default" language, and any untranslated messages will appear in the default language. For Cornish and Manx, English is fine as a default, but the Breton and Haitian Creole translators chose French as a default, Chechen uses Russian, Aragonese uses Spanish, etc.

Again with the aim of drawing new language groups into the world of software translation, we have encouraged translations of even a few key words, for example "Like" and "Unlike". Even this can have a lot of symbolic value for a language that is rarely seen on the computer, and is better than using the site entirely in English in any case.

The translation process itself is quite simple. Translations are stored in a plain text file that uses the standard file format used in most open source software translation (a so-called "PO file"). For users with some software translation experience, there are good tools for dealing with PO files, managing translation memories, terminology, etc. For others, it is easy enough to enter the translations with a simple text editor, and I am even willing to paste the translations into the correct format if someone emails them to me.

Here is an excerpt from one of the Facebook translation files (for the Chichewa language, spoken in Malawi, and translated by Edmond Kachale):

```
#. Button label at top of Groups page (click "MORE" next to "GROUPS")
msgid "Create Group"
msgstr "Pangani Gulu"

#. Link at bottom of home page (under "More") to info for software developers
msgid "Developers"
msgstr "Amisiri"

#. %a1 is a name, %a2 is a name or else something like "6 others"
msgid "%a1 and %a2 like this."
msgstr "%a1 ndi %a2 akonda ichi."
```

The English originals are on the lines marked “msgid”, and the translations are filled in on the lines marked “msgstr”. Note that the third example contains “variables” which are filled in with names, or translations of other messages from elsewhere in the PO file. More on these variables, and the linguistic mischief they can cause, below.

A major goal was to make the experience of using these translated versions as good (or better) than using any of the official Facebook translations. First and foremost, the script should not break any functionality. It should work on any platform, and with any web browser. And Facebook should not run any more slowly with the Greasemonkey script installed than it normally would. For the most part, we have achieved these aims, although in terms of browser support we generally recommend that the scripts be used with Google Chrome since the installation process is the simplest, and the scripts run somewhat faster with Chrome than they do with Firefox. In terms of linguistic quality, we have built some flexibility into our system that makes it possible to do *better* translations than the official Facebook ones, as discussed in the final section.

We are thrilled by how many language groups around the world have embraced this approach and are choosing to use Facebook in their own language. To date, 29 languages have submitted at least partial translations that are being used live on the site, and another 25 languages groups have begun work. Of these 54 languages, I am aware of previous software translation efforts in only 20, and 35 of the 54 appear in the UNESCO Atlas: 8 vulnerable languages, 13 definitely endangered, 7 severely endangered, and 7 critically endangered. Three others are not listed in the Atlas but probably should be: Silesian (szl) and Kven Finnish (fkv) are relatively new additions to the ISO 639-3 standard, and Powhatan is not listed at all since it is a “sleeping language”, though it is being reawakened through the efforts of Ian Custalow and other members of the Mattaponi tribe (Rising Voices, 2012).

Terminology wiki

Again, one of the aims of this project was to make it easier for language groups who have never translated software before to undertake a translation. A lack of technical terminology is a widespread problem for indigenous languages around the world. The Facebook interface has its share of technical terms, and Facebook has its own site-specific jargon as well: “status update”, “to like/unlike”, “to poke”, etc. There are also Western concepts like “Privacy” and “Advertising” that have been difficult to render into some indigenous languages.

A useful technique in terminology creation is to see how other languages have dealt with a given concept. If you speak French, Spanish, German, etc. it is easy enough to consult those translations directly for ideas. But someone translating into, say, an indigenous African language is less likely to be able to benefit from existing translations into a Native American language, for example. So I asked each translator to provide “back translations” into English of any of their translations that might be useful to other groups. These are stored in a Wiki on the project web site (Scannell et al, 2012) for easy updating and consultation by translators.

Here are some examples:

Activity Log

- French: “Personal history”
- Gundjeihmi: “What have you been doing”
- Irish: “Activity history”
- Nawat: “See what has happened”
- Spanish: “Record of activity”
- Chichewa: “Frequent activities”

Chat

- French: “Instant Discussion”
- Gundjeihmi: “Send each other talk”
- Irish: “Conversation”
- Ligurian: “Gossip”
- Chichewa: “Talks”

See Friendship

- French: “See the links of friendship”
- Ligurian: “Friendship details”
- Nawat: “What they have chatted about”
- Scots Gaelic: “the connection between you”

Unlike

- French/Ligurian: “I don’t like anymore”
- Hiligaynon: “Oh, this is not nice after all”
- Irish: “I don’t like it”
- Scots Gaelic: “Not like anymore”
- Spanish: “I don’t like”

When consulting these back translations, sometimes a certain phrasing or metaphor just “clicks” and can be carried over into another language. It can also help in clarifying the meaning of some of the more obscure messages; e.g. “Change Cover” which is rendered much more clearly, I think, in Nawat, as “Change the big picture”. Overall, I believe this is a great way for different endangered language groups to share metaphors and different “ways of knowing” to create terminology without always borrowing directly from English.

Linguistic details

In software translation, the translator is often constrained by choices made by the software developers, who are often unaware of the linguistic complexities that arise in many languages, or if they are aware, are unwilling to add special code to deal with these. There are a number of problems in the official Facebook translations that we have been able to solve in our unofficial translations.

One of the most common issues in software translation is plural-handling. To translate a message like “You have N new message(s)”, most European languages require two cases, like English: one for the singular and one for the plural. But in many other languages, the form the noun takes following a numeral N can depend in complicated ways on the value of N. In Irish, for “N things” we would say “2 cheann”, “3 cinn”, “7 gcinn”. The official Facebook translations into widespread languages like English, French, Spanish, and Russian seem to handle this correctly, but it is badly broken for the smaller languages I have tested like Irish and Welsh.

All of our unofficial translations, on the other hand, deal with plural forms correctly. The rules for many languages can be found in the Common Locale Data Repository (CLDR) maintained by the Unicode Consortium (2012), and the PO file format we use has a special syntax that supports separate translations depending on the value of a numeral N. Here is an example of this in the case of Scottish Gaelic:

```
msgid "1 person"
msgid_plural "%d people"
msgstr[0] "%d duine"
msgstr[1] "%d dhuine"
msgstr[2] "%d daoine"
msgstr[3] "%d duine"
```

The first translation is used when N=1 or 11, the second for N=2 or 12, the third for N=3,...19, and the last for all other cases.

The second major linguistic issue arises because of “translation templates” (sometimes called “sentence puzzles” by frustrated translators) like these:

- John Doe commented on a X.
- John Doe commented on your X.
- John Doe likes a X.
- John Doe shared Jane Doe’s X.

where in each case the variable X can be filled in with translations of one of a number of choices: “status”, “post”, “photo”, “link”, etc. The problem, of course, is that the translations of these individual words in many languages will depend on the surrounding context in which it is to be substituted. We give several examples below. This is not, as far as I know, a problem for the official Facebook translation into any major language, but is again badly broken in the official Irish, Basque, and Swahili translations, and perhaps others as well. One solution would be to require translators to provide separate translations of every possible combination of template and noun, but this quickly becomes unwieldy and would violate our commitment to keeping the size of the translation project capped at 200 messages. So instead, we added some flexibility to our translation system beyond that which is available in Facebook’s, allowing the translator to provide separate translations for the words “status”, “post”, etc. if the surrounding context warrants.

Noun case in Mwotlap. In Alexandre François’ Mwotlap translation, the noun X is translated using the locative case in sentences such as “John Doe likes a X” or “John Doe commented on a X”. But in a small number of templates, for instance “John Doe shared Jane Doe’s X”, the accusative case is required.

Celtic initial mutations. In all six Celtic languages, nouns frequently undergo initial mutations depending on the surrounding context, usually the preceding word. This leads to some of the most common (and egregious) errors in the official Irish translation. In Scottish Gaelic, *ceangal* is the usual translation of “link”, and so “John Doe likes a link” would be *'S toil le John Doe ceangal*. But if John Doe comments on your link, it becomes:

Thug	John Doe	seachad	beachd	air	a’	cheangal	aga
give-PAST-IND	John Doe	past	opinion	on	ART	$^{\text{LEN}}$link	at-2Sing

with an initial mutation (lenition) and the definite article. Thanks to Michael Bauer for providing this example.

Bantu noun classes. We will illustrate this issue with one simple case. In Edmond Kachale’s Chichewa translation, the templates containing a possessive (“John Doe’s X”) are translated as “the X of John Doe”. But the preposition corresponding to “of” needs to agree with the noun class of X, and consequently it cannot be included in the translation of the template, and instead must be bound to the translation of X itself. So *ulalo* (class 14, “a link”) and *chithunzi* (class 7, “a photo”) need to be rendered as *ulalo wa* and *chithunzi cha*, respectively, when substituted into possessive sentences.

Acknowledgements

I am grateful to everyone who has breathed life into this project by contributing translations into their native language and sharing their linguistic expertise: David Ojara Prince, Peter Rohloff, Bulat Betalgiry, Islam Elsanov, Sarah Slye, Joshua Verano, Federico L. G. Faroldi, Tony Scott Warren, Geraint Jennings, Michael Bauer, Dean Yibarbuk, Andrew Manakgu, Violet Lawson, Murray Garde, Chris Sheard, Adrian Cain, Francis Dimzon, Eliodora Dimzon, Edgar Siscar, Ofelia Libo-on-Salaya, Blomar Rain Catipunan, Essan Labos, Emmanuel Lerona, Ruben Magan Gamala, Jean Came Poulard, Steve Harris, Alessio Gastaldi, Karaitiana Taiuru, Alex François, José Pedro Ferreira, Edmond Kachale, Alan R. King, Greg Dickson, Gabe Archie, Carl Archie, Mohomodou Houssouba, and Jordan Kutzik. This paper is dedicated to the memory of Neskie Manuel (1980-2011).

References

Greasespot (2012). Greasemonkey: Grease up the series of tubes. Retrieved 14 July 2012, from: http://www.greasespot.net/

Haddad, G. (2009). Facebook now available in Arabic and Hebrew. Retrieved 11 July 2012, from: http://blog.facebook.com/blog.php?post=59043607130

Haddad, G. (2010). Facebook Translations & the Social Web. Retrieved 11 July 2012, from: http://www.w3.org/International/multilingualweb/madrid/slides/haddad.pdf

Little, C. (2008). Facebook in Translation. Retrieved 11 July 2012, from: http://blog.facebook.com/blog.php?post=20734392130

Manuel, N. (2010). Secwepemc Facebook: A Greasemonkey Userscript to alter Facebook to display words in Secwepemctsin. Retrieved 11 July 2012, from: https://github.com/neskie/secwepemc-facebook/blob/master/README.textile

Rising Voices (2012). Powhatan Language Revitalization. Retrieved 14 July 2012, from: http://rising.globalvoicesonline.org/grantees/powhatan-language-revitalization/

Scannell, K. (2011). Welcome/Fáilte! Retrieved 13 July 2012, from: http://indigenoustweets.blogspot.com/2011/03/welcomefailte.html

Scannell, K. et al (2012). Secwepemc-facebook wiki: Back Translations. Retrieved 14 July 2012, from: https://github.com/kscanne/secwepemc-facebook/wiki/Back-Translations

Unicode Consortium (2012). CLDR - Unicode Common Locale Data Repository. Retrieved 14 July 2012, from: http://cldr.unicode.org/