

Minority Report: Helping Less-resourced Languages to Share Data

Andrew Joscelyne
TAUS

The recent mobilization of translation technology and human resources to help translate in post-quake Haiti is a timely reminder of the problem facing resource-scarce languages in an age of expanding translation automation.

When the web emerged as the natural platform for communicating, trading and storing data in the 1990s, many thought that the language playing field would finally be leveled. Anyone could in principle acquire linguistic real estate on the web once a few coding standards had been agreed on. This would help sustain any of the 6,000 tongues extant on planet Earth today, whether they had 80 or 800 million speakers.

Yet minority – “small” or less-used – languages still have to fight for recognition in a world of big strategic languages, especially when it comes to digital language resources.

Being literate

Data-driven translation technology favors large communities of speakers (more specifically writers/readers) who generate and recycle text by the bucketful and translate the text into or from other big languages. Minority tongues are precisely those with small numbers of literates generating fewer usable language resources.

However, this classification would put languages such as Thai, Hindi or Bahasa Indonesian in the “minority” camp (that is, a relative scarcity of digital resources), even though they have a multimillion-strong speaker/writer base. One of Asia Online’s missions, as originally expressed by Dion Wiggins, is precisely to transform nondigital resources in Asia into data for translation automation.

Specialists tend to paraphrase minority by “marginalized,” especially in Europe where the newer big languages have constantly controlled communication from the economic center, relegating older smaller languages such as Basque, Breton, Catalan, Cornish, Corsica, Friul, Frisian, Galician, Occitan, Sami, Scottish Gaelic or Welsh to the periphery of national conversations, along with more recent arrivals such as Amharic, Bengali, Berber, Gujarati, Kurdish or Maghrebi Arabic.

At the same time, languages with relatively few speakers (well under a million) can act “big” as national languages

– witness Estonian, Icelandic, Irish and Maltese – and in this sense are only marginalized at a transnational European level.

For the burgeoning language resource sector, though, the real size of a literate population – and hence a resource base – is critical, mainly because it can be expensive to generate enough data to leverage the power of machines and automate translation at realistic prices. Some form of data sharing will be needed to pull these populations fully into the global linguasphere. The question is “Under which auspices do you share data?”

Sharing Welsh

Some minority languages play a pivotal role in public life (administration, tourism and so on) and develop a wide range of usable monolingual and possibly bilingual resources; others only have a domain-specific text such as the translated Bible as their written resource. And some have neither. There are efforts underway to improve the situation on the ground.

In an exemplary move, the Welsh Language Board (WLB) recently made available free of charge the rights to a TMX-based translation memory (TM) of various content from its website along with other core documents in Welsh. This TM is available on the WLB website, via the TAUS Data Association (www.tausdata.org), and also on Google.

This forms part of a more comprehensive strategy to make any resource found or created by the WLB available for reuse by the community. This is not strictly speaking an “open-source” approach, but does offer a way of recycling content originally paid for by the taxpayer. The European Commission and similar public bodies are also putting translated content back into the communities that helped generate it.



According to Jeremy Evas, leader of research, grants and language technology at the WLB, the aim is also to try and help normalize the Welsh language by circulating “approved” linguistic knowledge. For example, translators can download various terminology collections, ranging from restaurant menus to human resources-speak. There is also a plan to set up a national terminology center.

More content will be made available in due course, including the interface for a library catalog, under an open-source license. Another key document collection that could provide resources for a number of translation and research tasks is the bilingual (translated) proceedings of the Welsh Parliament. In general, Wales wants this content to circulate freely without licenses or restrictions of any kind, although certain academic linguists are not convinced that they will be quite as freely available as claimed.

Rather than set up a “national” Welsh project to build a machine translation (MT) system, the WLB is standing back and letting MT engines bloom freely. But it does supervise a rolling grant provision of some nine million euros a year for funding Welsh language technology innovations.

Back to Basque

In contrast, the Basque language community – another well-documented minority case that forms one of the many tongues in Spain’s linguistic quilt – has invested deeply in an open MT solution for translation between Spanish and Basque. It forms part of Spain’s national OpenTrad program launched several years ago that also stimulated the development of several (cognate) language pairs. The Basque part of the project developed a Castilian-Basque engine – Matxin – that now includes language pairs such as English and French in its ecosystem.

Matxin technology continues to be used by Eleka Ingeniaritza Linguistikoa, a Basque language technology company, to develop translation engines, and the firm is also looking at other open solutions such as Moses for certain languages and domains. But manager Inaki Irazabalbeitia Fernandez finds that “data is a big problem for translation development. We find that public administrations tend to hold on to their own data rather than share it.”

Eleka’s parent company, Elhuyar, has a foundation that finances language research and development projects that try to promote more data sharing on an international scale. Fernandez himself recently proposed to the policy principal for Basque language affairs that a public bank of Basque-Spanish and other language pairs should be set up using the extensive collection of government TMs.

Given the classic problems surrounding the “openness” of data acquisition, one of the most useful resources, at

least for technology developers in search of data, has been the collection of multilingual versions of Wikipedia and similar large-scale “crowdsourced” textual projects. The shortest path to developing usable resources for any language these days must surely be to translate/author a set of descriptive articles about the world using the Wikipedia format. In the pre-digital past, the primary source in question might have been the Bible and other (usually Christian) religious texts, as is still the case for Haitian language resources today.

Towards the metaphrasome

Indeed, translation automation folk should remember that back in the 1970s, well before most of our current tools and data collections existed, the Mormon Church in Utah planned to develop new IT technology to help them translate the Book of Mormon either in part or wholly into many hundreds of languages, most of them what we might call “minority” tongues.

In its initial phase, that research effort gave rise to the Weidner Communications’ language technology platform, a “multilingual word processing station” that was the remote ancestor of both SDL’s current MT system and the ALPS translation “memory” system, the granddaddy of today’s TMs.

Let’s hope that such social, religious, administrative or literary needs will drive more minority language communities to collect and curate a variety of multilingual language data and contribute them to what (inspired by the concept of the genome) we might start calling the metaphrasome – that is, the totality of translational possibilities, setting our sights not just on tens but hundreds of individual languages and thousands of language pairs.

Some form of data sharing will be needed to pull these populations fully into the global linguasphere.



TAUS is a think tank for the translation industry, undertaking research for buyers and providers of translation services and technologies.

Our mission is to increase the size and significance of the translation industry to help the world communicate better.

To meet this ongoing goal, TAUS supports entrepreneurs and principals in the translation industry to share and define new strategies through a comprehensive program of events, publications and communications.

Contact:

www.translationautomation.com

info@translationautomation.com