Digital Encoding of Small Minority Languages

Gabe DeFreitas

December 5, 2018

Contents

1 Introduction

Language is fundamentally a spoken phenomenon, but ever since the dawn of writing, the specter of written form has pulled a psychological trick over viewers. We view it as “the real thing” because we can hold it in our hands, roll it around, and sniff the paper it’s written on. We view it as “better”, since its author spent painstaking hours beneath the candlelight, honing each sentence to discover how far they might drive the limits of their craft. I’m looking at you, “Aliens For Dinner” [5]. And let’s not forget about other classic hits like “The Bible”, “The Qur’an”, and “The Corpus Juris Civilis”; the centrality of written texts to law and religion underpins the prestige of writing.

But in many ways, writing is the province of the few. Before widely available education in developed countries, literacy was a clear divider between haves and have-nots. Although literacy is widespread in the developed world today, the use of writing *on a large scale* (ie. publication) remains out of reach for many people. Powerful groups control traditional means of publication, such as publishing houses and news outlets. Today the Internet is changing this picture somewhat. Internet fora like blogs, social media sites, and websites are a low-cost way of submitting work to the wider world. Although being discovered outside your own network is difficult on Internet platforms, hosting sites at least give individuals' work a venue and a chance of being accessed by others.

The Internet introduces equal access problems of its own. Along with levelling access to media, the Internet is simultaneously levelling the language in which we access it. Due to the early development of computers and networking in the United States and to the economic power of English, English gained a tangible head start as the language of the Web. This is rapidly changing in the case of major languages, like Spanish, French, and Chinese as Unicode-based technologies make representing a wide array of languages much easier than in the past.

The real linguistic losers in the Internet game are small minority languages. By this term, we do not refer to isolated or nonstandard spoken languages or dialects. We are interested in languages with a written tradition, but which are relatively marginal to the state in which they are used. Examples might include langauges like Galician, Sindhi, Assamese, and Tamazight. These languages have not benefitted from state resources in developing digital linguistic infrastructure to non-English languages and thus find themselves even further behind.

First we will outline the goals and rationale of Unicode and UTF-8 for creating a multilingual Internet. Then we will see how Unicode adoption at the state level is the best means of promoting linguistic diversity on the web and protecting small languages by looking at cases from Vietnam and Myanmar.

2 What is Unicode

The unfortunate truth is that your computer can only read, write, or think in 1s and 0s. That mean any text a computer deals with must somehow be stored as a number; encoding schemes describe how the computer should convert this number into a string of characters or vice versa. All computers on a network must adhere to the *same* encoding system, lest streams of text be interpreted incorrectly and communication failures ensue.

The American Standard Code for Information Interchange (ASCII), a 7-bit encoding scheme, was released in 1963 and quickly became the worldwide standard for digital text encoding. It was even endorsed by President Lyndon Johnson, who ordered in 1968 that all federal government computers must be ASCII-compatible. This was a easy decision for the US President, as the 128 (2^7) characters in ASCII included all necessary symbols to represent modern English text. [6]

As soon as computer technology spread beyond the labs of Silicon Valley and government offices in Washington problems arose in adapting the technology to languages besides English. In other countries and languages, there was a vast array of linguistic forms, including diacritical marks, right-to-left scripts, and pictographic writing systems, which ASCII was unequipped to handle. In the 1990s, many national encoding schemes were created to represent languages other than English. Most of these were 8-bit "extensions" of ASCII of 256 characters (2^8) , meaning the original 128-character ASCII set remained intact, and useful non-English characters were added in the remaining slots. The only problem here is that they were not intercompatible, meaning that documents might not "play nice" when transferred between countries or languages. [6]

Unicode and its dominant encoding scheme UTF-8 intend to alleviate these problems by collecting all of the world's glyphs into one system. The 127 characters in ASCII are represented using the same single-byte codes, which means that no conversion for existing ASCII files is necessary, and these will not become bloated in a new conversion scheme. But the encoding scheme takes advantage of modern computer's greater storage capacities to represent up to 1,112,064 distinct characters.

The Unicode Consortium's desire is to map every conceivable linguistic symbol in the world to a unique digital representation. Unicode 11.0, released in June 2018, spans a grand total of 137,374 unique characters. Among this year's additions were support for several new scripts, such as Dogra, Gunjala Gondi, Medefaidrin, and Old Sogdian; 66 new emoji characters (sorry but I had to go there); Xiangqi Chinese chess symbols; and long-awaited support for "historic documents of the Buryats of the Barguzin Steppe" [7]. With a dizzying array of linguistic diversity represented, Unicode implementation is the clear choice today for a multilingual digital infrastructure.

3 Two Unicode Stories

3.1 Vietnam

The chapter by My, Huy, and Vilavong [3] discusses problems with representing Vietnamese minority languages in digital form. See the following key passage:

> In Vietnam, processing the Vietnamese-Kinh language problems has deployed fairly soon, had many results, and has been continued. However, the script problem on the computer for the ethnic minority languages has not been interested much. Especially, in the explosive development of information and communication technologies as well as internet, the services on the internet have been relatively familiar to the people in almost all regions of the country. **However, there is not any website in ethnic minority languages.** Even in the website of the Committee for Ethnic Minorities Vietnamese CEMA, there is not any ethnic minority language, the websites of the locals where the ethnic people live are only in

> Vietnamese-Kinh language, or accompanied by English.
> (My, Huy, and Vilavong [3])

Implementing digital resources in the Vietnamese state language proceeded quickly due to adequate resources. At the time of the paper, although Unicode was an established standard for national languages like Vietnamese, as Vietnamese typing tools in Unicode had been developed around the turn of the century [4]. Support lagged behind, however, for Vietnam's minority languages like Ede; thus it was necessary to implement an elaborate conversion scheme to represent Ede-specific characters. Ede is written using Vietnamese Latin characters, but uses a few letters not present in Vietnamese. In essence, the authors created an *ad hoc* encoding scheme for the language in their Microsoft Word plugin.

For the authors, the crucial aspect of Unicode is portability and intercompatibility. Unicode provides an extensible framework; more languages may be easily incorporated as standards developed. So instead of the situation described by the authors, in which "the script processing of the ethnic minority on the computer has only been solved locally for each ethnic language, have not been a national unity and have not satisfied the needs of the culture development and integration of ethnic minority communities in Vietnam," implementing Unicode solutions for these languages will allow written regional languages to operate seamlessly amongst each other and with the national language, Vietnamese.

3.2 Myanmar

Nearby in Myanmar, a very different story occurred. One early font and encoding package for Burmese, Zawgyi, gained early adoption in the country, based on the earlier font Myazedi. Burmese script, like other Brahmic abugidas, contains base letters with modifications or additions made around the base to signify vowels and other features. Although Unicode standard specifies that such features should be encoded separately and intelligently rendered into the correct output character, this took a while to be widely supported. Myazedi instead encoded each form separately resulting in a confusing situation:

> In Zawgyi, there are six different ways to write the word "myo" that render a superficially "correct" character, and many more if you allow for "incorrect" variations that would look strange but still intelligible to a reader. A computer, however, sees these variations as completely different words. Modern Unicode, by contrast, has only one code point per element, and will only render if the characters are encoded in the correct sequence, meaning that for each word there is one and only one encoding. (Hotchkiss [1])

The other significant problem is that the nonstandard encoding in Myazedi/Zawgyi used code space that had been reserved for Myanmar's small languages, like Shan, Mon, Kayah, and Karen.

For legacy reasons, however, Zawgyi remained the dominant encoding in the country at the time of the article. Smartphones by Samsung and Huawei ship with Zawgyi as default, and the popular web portal planet.com.mm, included instructions on how to install Zawgyi on your system. Myanmar blogs also offered guides for setting up Zawgyi [1]. It was seen as a homegrown, familiar font, in contrast to difficult and foreign Unicode fonts. Lack of Unicode adoption, driven by lack of information among common users, is limiting Myanmar's participation in modern digital culture and its representation of all the linguistic cultures present within its borders: "New users in Myanmar unknowingly become part of Zawgyi's existing user base, without knowing the hidden costs that will impede the future of Myanmar's digital society to be sustainable and inclusive." [2]

References

[1] Hotchkiss, Griffin. *The complex battle over Burmese fonts, explained*. 2016. URL: `https://coconuts.co/yangon/news/complex-battle-over-burmese-fonts-explained/`.

[2] Liao, Han-Teng. "Encoding for Access: How Zawgyi success impedes full participation in digital Myanmar". In: *ACM Computers and Society* 46.4 (2017).

[3] My, Le Hoang Thi, Huy, Khanh Phan, and Vilavong, Souksan. "Using Unicode in Encoding the Vietnamese Ethnic Minority Languages, Applying for the Ede Language". In: (2013).

[4] Nguyen, Thi. "Typing Vietnamese, Part 2: The Vietnamese Diaspora, Unicode and the Ubiquity of Unikey". In: *Saigoneer* (2018). URL: `https://saigoneer.com/saigon-technology/14055-typing-vietnamese,-part-2-the-vietnamese-diaspora,-unicode-and-the-ubiquity-of-unikey`.

[5] Spinner, Stephanie and Björkman, Steve. *Aliens for Dinner*. New York: Random House, 1994.

[6] Tero, Paul. *Unicode, UTF8 and Character Sets: The Ultimate Guide*. 2012. URL: `https://www.smashingmagazine.com/2012/06/all-about-unicode-utf8-character-sets/`.

[7] *Unicode® 11.0.0*. Unicode. 2018. URL: `http://unicode.org/versions/Unicode11.0.0/`.