Digitally Encoding Small Languages

Gabe DeFreitas

December 5, 2018

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” [9]. And let’s not forget about other classic hits like “The Bible”, “The Qur’an”, and “The Corpus Juris Civilis”. In all seriousness, the central position of written texts in law and religion underpins the prestige of writing.

In many ways, writing is the province of the “few”, a tool for the upper classes. Although this situation has equalized in developed countries somewhat given increased literacy rates, (ie. the interaction with and production of written media are an essential aspect of modern life) it remains true today that harnessing writing *on the large scale* (publication) remains out of reach for most people. Powerful organizations control traditional means of publication, including publishing houses, news outlets, and governments. The Internet is changing the situation somewhat. Internet fora like blogs, social media sites, and web pages are a low-cost way of presenting work to the wider world. Although having this work viewed outside your own network is still difficult on Internet platforms, hosting sites at the least give individuals’ work a venue and a chance of being accessed by others.

But then the Internet introduces equal access problems of its own. For one, there is the obvious issue of global disparities regarding access to technology and data. These disparities exist between developed and developing countries, as well as among different classes within the same country. Alternatively the Internet may be levelling the language in which we access written media. 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 most recent study by Fundación Redes y Desarollo found 32% of Internet content to be written in English, followed by 18.0% in Chinese, 8.0% in Spanish, and 6.5% in French [5].

The real linguistic losers on the Internet are small minority languages. By this term, I do not mean extremely isolated or nonstandard spoken languages. Instead, I am referring to languages with a written tradition, but which are not widely used in official state-level contexts. Examples include languages like Galician, Sindhi, Assamese, Tamazight, and Sardinian. These languages do not benefit from state resources for developing digital linguistic infrastructure. Likewise, speakers of regional languages, especially in developing countries, may be negatively impacted by the global digital divide. Finally, there is the ever-present threat that English and other large national languages will devour the world's local languages.

First I will briefly discuss the rationale and goals of Unicode for building multilingual digital architecture. Then we will look at two case studies (Vietnam and Myanmar) to see how the choice of encoding standards can support or discourage the digital growth for languages. In Vietnam, Unicode enjoys wide support; this makes Vietnamese documents compatible with the wider world and provides an extensible framework for the inclusion of Vietnam's ethnic languages. In Myanmar, however, Unicode support lags behind due to the popularity of a legacy encoding called Zawgyi; this situation threatens Myanmar's compatibility with modern software and the representation of Myanmar's ethnic languages.

2 What is Unicode

Unfortunately your computer can only read, write, or think in 1s and 0s. That means any text your computer encounters must somehow be stored as a number; encoding schemes are standards that describe how the computer should convert a 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 breakdowns ensue. (The Japanese term mojibake (文字化け) refers to "the garbled text that is the result of text being decoded using an unintended character encoding" [6].)

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. [10]

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. [10]

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" [11]. With a dizzying array of linguistic diversity represented, Unicode implementation is the clear choice today for a multilingual digital infrastructure.

3 Two Unicode Stories in Southeast Asia

3.1 Vietnam

My, Huy, and Vilavong [7] considers problems and solutions to digital representation of Vietnamese minority languages, specifically Ede (Rade). 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 commu-

> nication 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 [7])

Digital resources for the Vietnamese state language developed relatively quickly due to economic incentives and state resources. (Vietnam's Ministry of Science, Technology, and the Environment adopted a Vietnamese 8-bit ASCII extension, Vietnamese Standard Code for Information Interchange (VSCII), in 1993 [2].) By the time of the paper, Unicode had replaced legacy standards and Vietnamese typing tools in Unicode had been developed [8]. Support lagged behind, however, for Vietnam's minority languages, like Ede; it was necessary to implement an elaborate conversion scheme to represent Ede-specific characters. Ede is written using Vietnamese Latin characters, but includes some letters not present in Vietnamese. (For example, the letter Ƀ used in Ede was not supported in Unicode until 2006, while Vietnamese typing tools like Unikey and Vietkey were developed prior to this inclusion. [3]) My, Huy, and Vilavong [7] essentially present an *ad hoc* encoding scheme for the language in their Microsoft Word plugin that adapts the Unicode-based typing tools to handle Ede characters.

The crucial benefit of Unicode in Vietnam is portability and intercompatibility. Unicode provides an extensible framework; more languages may be easily incorporated as standards developed. The limitation of Vietnamese typing tools is due to lack of upgrades to their software, rather than to lack of ambition by the Unicode Consortium for supporting minority languages. As My, Huy, and Vilavong [7] state: "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

In nearby Myanmar, a very different story played out. One early font and encoding package for Burmese, Zawgyi, based on the earlier font Myazedi, gained early market dominance in the country. Burmese script, like other Brahmic abugidas (Devanagari, Tamil, etc.) consists of base letters with modifications and diacritics inserted around the base to mark vowels and other features. Although Unicode specifies that such features should be encoded separately and intelligently combined by software into the correct output character, advanced rendering took a long time to be widely supported. Zawgyi's predecessor Myazedi instead encoded each combined form in its own code point, 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. ([1])

Lack of adherence to Unicode standards had another unintended effect. The nonstandard encoding pattern used by Myazedi/Zawgyi took up code space that had been reserved for Myanmar's ethnic languages, like Shan, Mon, Kayah, and Karen [1]. Thus Zawgyi excludes from the outset the inclusion of Myanmar ethnic identities in the country's digital infrastructure.

For legacy reasons, Zawgyi remained the dominant encoding in the country at the time of the article in 2016. By this time, in contrast, Unicode and UTF-8 had become nearly ubiquitous throughout much of the world. Smartphones by Samsung and Huawei ship with Zawgyi as default, while the popular web portal planet.com.mm, includes 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 is driven by lack of information among average users, thus limiting Myanmar's full participation in global digital infrastructure and the representation of all linguistic traditions 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." [4]

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] jdo@emperor.mentorg.com. *Re: CRL Archive Change Announcement*. 1993. URL: `https://www.informatik.uni-leipzig.de/~duc/software/misc/tcvn.txt`.

[3] "LATIN CAPITAL LETTER B WITH STROKE". In: *English Wikipedia* (2018). URL: `https://en.wikipedia.org/wiki/%C9%83`.

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

[5] *Mesure des langues dans l’Internet (2017)*. Fundación Redes y Desarollo. 2017. URL: http://funredes.org/lc2017/.

[6] “Mojibake”. In: *English Wikipedia* (2018). URL: https://en.wikipedia.org/wiki/Mojibake.

[7] 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: Firth International Conference on Knowledge and Systems Engineering. Ed. by Van-Nam Huynh et al. Hanoi, 2013.

[8] 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.

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

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

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