I Can Text You A Pile of Poo, But I Can’t Write My Name

We can’t ignore the composition of the Unicode Consortium’s members, directors, and officers -- the people who define the everyday writing systems of all languages across the globe.
by Aditya Mukerjee on March 17th, 2015

I am an engineer, and I am a writer. As an engineer, I spend a lot of time thinking about how text is stored, but relatively little about what information the text actually represents. To the computer, text is an abstract entity — a stream of 0s and 1s, and any semantic meaning is in the eye of the beholder. As a writer, of course, the meaning is everything, and the mechanics of how the text is stored is merely a technical detail.

But in an economy that is increasingly digital, increasingly global, and increasingly multilingual, we can no longer maintain this distinction. The information we want to represent is intimately linked to how it is stored. We can no longer separate the two.

What is Unicode?

The globe, with the North American continent outlined in binary code.

Written text is a sequence of graphemes — characters. Every character you type — whether letters like ‘a’ and ‘b’, punctuation marks like ‘?’, or even emoji like 💪👬👍 – has an ID number that computers use to store it. In order to communicate, computers need to agree on a common roster of how to assign these graphemes to numbers and vice versa. These rosters are known as character encodings. If you’ve ever stumbled across a website that looked readable, but certain characters were printed incorrectly (like “won’t” showing up as “wonâ€™t”), that’s because the computers were using slightly different encodings (in this case, UTF-8 and ISO-8859-1).

Until 2007, the most popular encoding was called ASCII. ASCII was first used in 1963, and later endorsed by President Lyndon B. Johnson. ASCII is formally known as US-ASCII, so it may come as no surprise that it focused exclusively on graphemes used in writing English. Today, the most common encoding on the Internet is UTF-8 (Unicode), which extends ASCII to support other languages.

Unicode was first conceived in the late 80s by Lee Collins, Joe Becker, and Mark Davis, engineers at Apple and Xerox, as an attempt to create a universal character set for all text, not just English text. Together, they founded the Unicode Consortium, which determines the list of Unicode characters and has published 18 versions of the list since 1991.

The founders and current members of the Unicode Consortium are well-credentialed. Collins has an M.A. from Columbia University in East Asian languages and cultures. Becker’s biography notes that he “speaks survival-level Chinese, French, German, Japanese, and Russian, and has forgotten

Latin.” (source)

Despite this, we can’t ignore the composition of the Consortium’s members, directors, and officers, the people who define the everyday writing systems of all languages across the globe. They are comprised largely of white men (and a few white women) whose first language was either English or another European language.

Many of them work for one of the nine organizations that hold full membership in the Consortium. Seven of these nine are US-based technology companies: Adobe, Apple, Google, IBM, Microsoft, Oracle, and Yahoo. One (SAP) is a German technology company. These companies have, by their very own admission, workforces that are overwhelmingly white, and leadership and tech teams that are even less represented by racial minorities. These reports lump Indians together under the broad umbrella “Asians”, so while it’s impossible to know exactly how badly various ethnic groups or native languages are actually represented, the data is far from encouraging.

The only other full member from the rest of the world is Oman’s Ministry of Endowment and Religious Affairs. Bizarrely, they are represented in the Consortium not by an Omani man, but by a Dutch man with only “professional working proficiency” in Arabic and a “full professional proficiency” in English, in addition to his native Dutch.

Second-Class Languages

My family’s native language, which I grew up speaking, is far from a niche language. Bengali is the seventh most common native language in the world, sitting ahead of the eighth (Russian) by a wide margin, with as many native speakers as French, German, and Italian combined.

And yet, on the Internet, Bengali is very much a second-class citizen — as are Arabic (#5), Hindi (#4), and Mandarin (#1) — any language which is not written with the Latin alphabet.

The very first version of the Unicode standard did include Bengali. However, it left out a number of important characters. Until 2005, Unicode did not have one of the characters in the Bengali word for “suddenly”. Instead, people who wanted to write this everyday word had to combine three separate, unrelated characters. For English-speaking teenagers, combining characters in unexpected ways, like writing ‘w’ as ‘\/\/’, used to be a way of asserting technical literacy through “l33tspeak” — a shibboleth for nerds that derives its name from the word “elite”. But Bengalis were forced to make similar orthographic contortions just to write a simple email: ত + ্ + ‍ = ৎ (the third character is the invisible “zero width joiner”).

Even today, I am forced to do this when writing my own name. My name is not only a common Indian name, but one of the top 1,000 names in the United States as well. But the final letter has still not been given its own Unicode character, so I have to use a substitute.

A few other characters that were more common historically, though still used today, were also missing for the first decade of Bengali’s existence in Unicode. It’s tempting to argue that historical characters have no place in a character set intended for computers. On the contrary, this makes their inclusion even more vital: rendering historical texts accurately is key to ensuring their survival in the transition to the age of digital media.

Furthermore, these characters are still common enough that they were printed in the Bengali reading textbooks and workbooks that we used growing up. Omitting them literally ensures that existing materials for learning to read Bengali will not be universally accessible.

Without that argument, one might appeal to the limited space in the Unicode character set. Even if we take for granted the somewhat arbitrary maximum of 1,114,112 codepoints, the other alphabets included speak for themselves. The most recent update to the Unicode standard included the entire alphabet of Linear B, an ancient Mycenaean script that was not deciphered in the modern era until the 1950s. Nor does alleged scarcity explain the inclusion of Linear A, a Minoan script so arcane, scholars disagree on what language it even represented, let alone how to read the script.

The East Asian Problem

I am not the only one who has trouble writing their name correctly in Unicode. Linguistically, East Asian languages such as Chinese, Japanese, and Korean have distinct writing systems. Some (but not all) of the characters trace their lineage back to a common set, but even these characters, known as Han characters, began to diverge and evolve independently over two thousand years ago.

The Unicode Consortium has launched a very controversial project known as Han Unification: an attempt to create a limited set of characters that will be shared by these so-called “CJK languages.” Instead of recognizing these languages as having their own writing systems that share some common ancestry, the Han unification process views them as mere variations on some “true” form.

To help English readers understand the absurdity of this premise, consider that the Latin alphabet (used by English) and the Cyrillic alphabet (used by Russian) are both derived from Greek. No native English speaker would ever think to try “Greco Unification” and consolidate the English, Russian, German, Swedish, Greek, and other European languages’ alphabets into a single alphabet. Even though many of the letters look similar to Latin characters used in English, nobody would try to use them interchangeably. ƫЋat ωoulδ βε σutrageϵθus.

Even though our language is exempt from this effort, Han unification is particularly troubling for Bengali speakers to hear about. The rhetoric is a blast from our own colonial past, when the British referred to Indian languages pejoratively as “dialects”. Depriving their colonial subjects of distinct linguistic identities was a key tactic in justifying their brutal rule over an “uncivilized” people.

This common sentiment is documented perfectly in My Fair Lady (based on George Bernard Shaw’s Pygmalion). Colonel Pickering proudly announces himself as ‘a student of Indian dialects’. Despite having just returned from studying Indian languages firsthand in India, Pickering has apparently failed to grasp the most basic of differences between India’s multitudinous languages. He proudly announces that there are ‘no fewer than 147 Indian dialects’ – a pathetically inaccurate count. (Today, India has 57 non-endangered and 172 endangered languages, each with multiple dialects – not even counting the many more that have died out in the century since My Fair Lady took place).

The Evolution of Emoji

At the same time, Unicode embraces a new pictorial writing system: emoji. There are almost 1,000 emoji characters, such as ❤ (“Heavy black heart”), but also 🎼 (“Musical score”), and even 📠 (“Fax machine”). Each emoji gets its own Unicode codepoint. They are varied enough in meaning that they almost comprise their own language, as evidenced by the recent translation of Melville’s classic Moby Dick into Emoji.

Maybe this isn’t so surprising; after all, various works of literature have already been translated into Klingon, an artificial language that only has three times as many words as emoji has characters. (Incidentally, in 2001 the Unicode Consortium rejected a 1997 proposal to include the Klingon alphabet known as pIqaD [sic].)

The evolution of emoji is impressive and fascinating, but it makes for an uncomfortable contrast when other pictorial writing systems – the most commonly-used writing systems on the planet – are on the chopping block. We have an unambiguous, cross-platform way to represent “PILE OF POO” (💩), while we’re still debating which of the 1.2 billion native Chinese speakers deserve to spell their own names correctly.

The Diversity Solution

Traditionally, many emoji fonts have used a bright yellow skin tone. To demonstrate its commitment to diversity in Unicode, the Unicode consortium recently announced a new proposal, which will allow users to specify any of six different shades of brown for emoji faces. If the right icon is available, it will be shown; otherwise, a “default” face will be shown, next to a color swatch (so you can make it explicitly clear to your friends that you were, in fact, thinking about race when you texted them “White Frowning Face” (☹), “Man with Turban” (👳), or “Man with Gua Pi Mao” (👲).

Perhaps I wouldn’t mind that the emoji world now literally has “colored” people, if it weren’t for the timing. Instead, what could have been a meaningless, empty gesture becomes an outright insult. You can’t write your name in your native language, but at least you can tweet your frustration with an emoji face that’s the same shade of brown as yours!

Writing for the 21st Century

Determining which graphemes and glyphs are essential to a given ethno-linguistic group is a tough problem. Identifying all of these for all languages in widespread use is even more challenging. But one thing is clear: we cannot design an alphabet meant for everyday use by native speakers of a language without the primary input of native speakers of these languages.

Out of compatibility concerns, the Unicode Consortium is unlikely to modify the 224,024 characters that have already been defined in any future updates. It took half a century to replace the English-only ASCII with Unicode, and even that was only made possible with an encoding that explicitly maintains compatibility with ASCII, allowing English speakers to continue ignoring other languages.

But that still leaves 80% of the codepoints unused. As the Unicode Consortium decides which characters to allocate, there are a number of ways to ensure that Unicode accurately reflects the stated goal of representing “all characters in widespread use today”.

Membership in the Consortium is not free, or even cheap. Full membership and voting rights cost $18,000 (and tellingly, all prices are listed in USD only). Discounts are already provided at lower membership tiers for non-profit organizations, such as the Mormon church. These discounts could be expanded to full membership, and to for-profit groups from non-European countries where English is a minority language. The Consortium could establish an explicit hiring plan to guarantee that its staff represent the many languages that it seeks to standardize. The Consortium could adopt bylaws that ensure that technical committee members and officers are not dominated by native English speakers. There are other measures that the Consortium can and should take as well, but these three are very straightforward both to implement and to evaluate, so they make a good starting point.

Gayatri Chakravorty Spivak has written, ‘The subaltern cannot speak’. They are structurally prohibited from having any dialogue – even an unbalanced one – with the very powers that oppress them. Access to digital tools that respect our languages is crucial to communicating in the Internet age. The power to control the written word is the ability both to amplify voices and to silence them. Anyone with this power must wield it with caution.

Whatever path we take, it’s imperative that the writing system of the 21st century be driven by the needs of the people using it. In the end, a non-native speaker – even one who is fluent in the language – cannot truly speak on behalf the monolingual, native speaker. For them, the language is simply a way of exploring a different part of their world, or of exploring familiar parts in a new way. For the native speaker, the language is not merely a novelty. It is the gateway to accessing life and society itself.