1 of 57 DOCUMENTS

The New York Times

October 22, 2017 Sunday
Late Edition - Final

The Content of Their Characters

BYLINE: By **MICHAEL ERARD**.

Michael Erard is writer in residence at the Max Planck Institute for Psycholinguistics and the author of "Babel No More: The Search for the World's Most Extraordinary Language Learners."

SECTION: Section MM; Column 0; Magazine Desk; FEATURE; Pg. 35

LENGTH: 2834 words

Anshuman Pandey was intrigued. A graduate student in history at the University of Michigan, he was searching online for forgotten alphabets of South Asia when an image of a mysterious writing system popped up. In eight years of digging through British colonial archives both real and digital, he has found almost 200 alphabets across Asia that were previously undescribed in the West, but this one, which he came across in early 2011, stumped him. Its sinuous letters, connected to one another in cursive fashion and sometimes bearing dots and slashes above or below, resembled those of Arabic.

Pandey eventually identified the script as an alphabet for Rohingya, the language spoken by the stateless and persecuted Muslim people whose greatest numbers live in western Myanmar, where they've been the victims of brutal ethnic cleansing. Pandey wasn't sure if the alphabet itself was in use anymore, until he lucked upon contemporary pictures of printed textbooks for children. That meant it wasn't a historical footnote; it was alive.

An email query from Pandey bounced from expert to expert until it landed with Muhammad Noor, a Rohingya activist and television host who was living in Malaysia. He told Pandey the short history of this alphabet, which was developed in the 1980s by a group of scholars that included a man named Mohammed Hanif. It spread slowly through the 1990s in handwritten, photocopied books. After 2001, thanks to two computer fonts designed by Noor, it became possible to type the script in word-processing programs. But no email, text messages or (later) tweets could be sent or received in it, no Google searches conducted in it. The Rohingya had no digital alphabet of their own through which they could connect with one another.

Billions of people around the world no longer face this plight. Whether on computers or smartphones, they can write as they write, expressing themselves in their own linguistic culture. What makes this possible is a 26-year-old international industrial standard for text data called the Unicode standard, which prescribes the digital letters, numbers and punctuation marks of more than 100 different writing systems: Greek, Cherokee, Arabic, Latin, Devanagari -- a world-spanning storehouse of languages. But the alphabet that Noor described wasn't among them, and neither are more than 100 other scripts, just over half of them historical and the rest alphabets that could still be used by as many as 400 million people today.

Now a computational linguist and motivated by a desire to put his historical knowledge to use, Pandey knows how to get obscure alphabets into the Unicode standard. Since 2005, he has done so for 19 writing systems (and he's currently working to add another eight). With Noor's help, and some financial support from a research center at the University of California, Berkeley, he drew up the basic set of letters and defined how they combine, what rules govern punctuation and whether spaces exist between words, then submitted a proposal to the Unicode Consortium, the organization that maintains the standards for digital scripts. In 2018, seven years after Pandey's discovery, what came to be called Hanifi Rohingya will be rolled out in Unicode's 11th version. The Rohingya will be able to communicate online with one another, using their own alphabet.

As a practical matter, this will not have much impact for the Rohingya who are suffering in Myanmar, many of whom are illiterate and shut off from educational and technological opportunity. "The spread of this new digital system is unlikely to go to scale," Maung Zarni, a human rights activist who works on Rohingya issues, and Natalie Brinham, a Ph.D. fellow at Queen Mary

University of London, told me in an email. They emphasized that the Rohingya do not have the autonomy to organize their own schools. But given the group's history of oppression, the encoding of their language carries considerable symbolic weight because it legitimizes an oppressed minority and their language. "It becomes a tool of unity to help people come together," Noor says.

Creating such interconnectedness and expanding the linguistic powers of technology users around the world is the whole point of Unicode. If the work is slow, that's because standardizing a writing system for computers is a delicate art that relies on both engineering and diplomacy. And the time and attention of the volunteers who maintain the standard are finite. So what happens when a new system of visual communication like emoji emerges and comes under their purview? Things get even slower and the mission more complicated.

Shortly after finishing a linguistics Ph.D. at Berkeley in 1980, Ken Whistler was frustrated by the inability of mainframe computers to print the specialized phonetic symbols that linguists use. I can fix that, he thought, and he then hacked an early personal computer to do so. In 1989, on one of his first days on the job at a software start-up, his boss told him to meet with a Xerox computer scientist, Joe Becker, who had just published a manifesto on multilingual computing. "The people of the world need to be able to communicate and compute in their own native languages," Becker wrote, "not just in English."

At the time, computing in the United States relied on encodings like those from the American Standard Code for Information Interchange (usually known as ASCII), which assigned numerical identifiers to letters, numbers, punctuation and behaviors (like "indent"). The capital letter "A," for instance, had an ASCII code of 065, or 01000001 in the 0s and 1s in the binary code of computers. Each textual character used by a computer needs its own unique sequence, a numerical identifier or "character encoding." The problem with ASCII was that it had only 256 codes to distribute and far more than 256 characters in the world needing identifiers.

In order to work with more writing systems than ASCII was able to handle, technology companies like Apple, Xerox, IBM, DEC, Hewlett-Packard and even Kodak created their proprietary encodings. None of them worked with the others. To complicate things further, some nations insisted as a matter of national pride on their own standards for text data. "The proliferation of character encodings was chaos," Whistler says.

Joe Becker gathered like-minded computer scientists to bring order to the chaos, arguing that cooperation was needed among companies. The result was the Unicode Consortium, which was incorporated in 1991. He also maintained that the solution had to be international and helped broker an alliance with the International Organization for Standardization, which maintains more than 20,000 standards related to products and services, from the tensile strength of yarn to the chemical composition of toys. Such standards are meant to ensure, among other things, that things from one country can be used in the industrial processes of another. Standardized shipping containers, for instance, have made international commerce far more efficient. Standards don't become regulations; they're conventions, "recipes for reality" in the words of Lawrence Busch, a sociologist emeritus at Michigan State University who studies how standards arise. Unicode unified all the numerical identifiers and made sure they were reliable and up-to-date.

As is the case in other standards organizations, full membership in the nonprofit Unicode Consortium comes with the right to vote on changes to the standard. Membership dues are $18,000 annually; current full members include global tech giants (like Apple, Facebook and Google) and the Sultanate of Oman (which wants to see digital Arabic improved). A second membership tier includes a university, government bodies in Bangladesh and India, a typeface company and an emoji search engine. Over the years, members came and went, depending on their immediate interest in issues of standardization.

Unicode's idealistic founders intended to bring the personal-computing revolution to everyone on the planet, regardless of language. "The people who really got the bug," Whistler says, "saw themselves at an inflection point in history and their chance to make a difference." No fortunes have been made through Unicode, unless you count the platforms (like Twitter) and products (like the iPhone) that adopted the standard.

Unicode's history is full of attacks by governments, activists and eccentrics. In the early 1990s, the Chinese government objected to the encoding of Tibetan. About five years ago, Hungarian nationalists tried to sabotage the encoding for Old Hungarian because they wanted it to be called "Szekley-Hungarian Rovas" instead. An encoding for an alphabet used to write Nepal Bhasa and Sanskrit was delayed a few years ago by ethnonationalists who mistrusted the proposal because they objected to the author's surname. Over and over, the Unicode Consortium has protected its standard from such political attacks.

The standard's effectiveness helped. "If standards work, they're invisible and can be ignored by the public," Busch says. Twenty years after its first version, Unicode had become the default text-data standard, adopted by device manufacturers and software companies all over the world. Each version of the standard ushered more users into a seamless digital world of text. "We used to ask ourselves, 'How many years do you think the consortium will need to be in place before we can publish the last version?' " Whistler recalls. The end was finally in sight -- at one point the consortium had barely more than 50 writing systems to add.

All that changed in October 2010, when that year's version of the Unicode standard included its first set of emojis.

On a downtown San Francisco street last November, partygoers were lined up at a Taco Bell truck to get tacos. Inside the nearby co-working space, Covo, was the opening night party of Emojicon, a weekend-long celebration of all things emoji, held just days before the presidential election. Only foods that could be depicted with emojis were being served, while a balloon artist twisted approximations of various emojis.

In the late 1990s, when Japanese phone manufacturers first put emojis on their devices as marketing gimmicks, messaging standards required that emojis be sent as text data -- as characters matched to strings of numbers, not as images. But emojis were unreadable on devices that couldn't translate their numerical identifiers.

When a software engineer named Graham Asher suggested in 2000 that Unicode take responsibility for emojis, the consortium demurred on the grounds that pictures were subjectively interpreted. A few years later, companies like Apple and Microsoft realized that the increasingly popular Japanese emojis would appear as gibberish on their products and pushed the consortium to encode them. By 2009, 974 emojis had been assigned numerical identifiers, which were released the following year.

As the demand for new emojis surged, so, too, did the criticisms. White human figures didn't reflect the diversity of real skin colors. Many emojis for specific professions (like police officer and construction worker) had only male figures, while icons for foods didn't represent what people around the world actually ate. Millions of users wanted to communicate using the language of emoji, and as consumers, they expected change to be swift. One thing appeared to be slowing things down: the Unicode Consortium.

At Emojicon, resentment toward Unicode was simmering amid the emoji karaoke, emoji improv and talks on emoji linguistics. "Such a 1980s sci-fi villain name," one participant grumbled. "Who put them in charge?" A student from Rice University, Mark Bramhill, complained that the requirements for the yoga-pose emoji he had proposed were off-puttingly specific, almost as if they were meant to deter him. A general antiestablishment frustration seemed to be directed at the ruling organization. One speaker, Latoya Peterson, the deputy editor of digital innovation for ESPN's "The Undefeated," urged people to submit proposals to Unicode for more diverse emojis. "We are the internet!" she said. "It is us!"

On the first morning of Emojicon, Mark Davis, president of Unicode, explained in a talk that the consortium also maintains the repository for time and date formats, currency and language names and other information that adapts computer functions to where a user is. Even more demanding technically is making sure that characters behave as users want them to. One major achievement has been ironing out how right-to-left alphabets like Arabic are used in the same line of text as left-to-right ones like Latin, which affects billions of users and can take years to adjust. Dealing with emojis, in short, is a small, though increasing, part of the consortium's responsibilities.

Davis mentioned that once characters become part of the Unicode standard, they're never removed. This inspired one young designer in the audience to announce that he'd ensure his legacy by proposing emojis until one was accepted. The crowd laughed; Davis smiled coolly, perhaps because Unicode committees have been overwhelmed with some 500 submissions in the last three years.

Not everyone thinks that Unicode should be in the emoji business at all. I met several people at Emojicon promoting apps that treat emojis like pictures, not text, and I heard an idea floated for a separate standards body for emojis run by people with nontechnical backgrounds. "Normal people can have an opinion about why there isn't a cupcake emoji," said Jennifer 8. Lee, an entrepreneur and a film producer whose advocacy on behalf of a dumpling emoji inspired her to organize Emojicon. The issue isn't space -- Unicode has about 800,000 unused numerical identifiers -- but about whose expertise and worldview shapes the standard and prioritizes its projects.

"Emoji has had a tendency to subtract attention from the other important things the consortium needs to be working on," Ken Whistler says. He believes that Unicode was right to take responsibility for emoji, because it has the technical expertise to deal with character chaos (and has dealt with it before). But emoji is an unwanted distraction. "We can spend hours arguing for an emoji for chopsticks, and then have nobody in the room pay any attention to details for what's required for Nepal, which the people in Nepal use to write their language. That's my main concern: emoji eats the attention span both in the committee and for key people with other responsibilities."

Emoji has nonetheless provided a boost to Unicode. Companies frequently used to implement partial versions of the standard, but the spread of emoji now forces them to adopt more complete versions of it. As a result, smartphones that can manage emoji will be more likely to have Hanifi Rohingya on them too. The stream of proposals also makes the standard seem alive, attracting new volunteers to Unicode's mission. It's not unusual for people who come to the organization through an interest in emoji to end up embracing its priorities. "Working on characters used in a small province of China, even if it's 20,000 people who are going to use it, that's a more important use of their time than deliberating over whether the hand of my yoga emoji is in the right position," Mark

Bramhill told me.

Since its creation was announced in 2015, the "Adopt a Character" program, through which individuals and organizations can sponsor any characters, including emojis, has raised more than $200,000. A percentage of the proceeds goes to support the Script Encoding Initiative, a research project based at Berkeley, which is headed by the linguistics researcher Deborah Anderson, who is devoted to making Unicode truly universal. One the consortium recently accepted is called Nyiakeng Puachue Hmong, devised for the Hmong language by a minister in California whose parishioners have been using it for more than 25 years. Still in the proposal stage is Tigalari, once used to write Sanskrit and other Indian languages.

One way to read the story of Unicode in the time of emoji is to see a privileged generation of tech consumers confronting the fact that they can't communicate in ways they want to on their devices: through emoji. They get involved in standards-making, which yields them some satisfaction but slows down the speed with which millions of others around the world get access to the most basic of online linguistic powers. "There are always winners and losers in standards," Lawrence Busch says. "You might want to say, ultimately we'd like everyone to win and nobody to lose too much, but we're stuck with the fact that we have to make decisions, and when we make them, those decisions are going to be less acceptable to some than to others."

URL: https://www.nytimes.com/2017/10/18/magazine/how-the-appetite-for-emojis-complicates-the-effort-to-standardize-the-worlds-alphabets.html

LOAD-DATE: October 22, 2017

LANGUAGE: ENGLISH

GRAPHIC: DRAWING (DRAWING BY MATT DORFMAN) (MM34-MM35)

PUBLICATION-TYPE: Newspaper

Print This Page Close Window