This paper describes Unicode, a worldwide standard that will remove most of the problems people have had in working with fonts and scripts for minority languages. The change is coming; it is unstoppable. You need to know what is going on.

Woe to those who use customized fonts!

Anybody who has used a computer for language work should be familiar with the problem of fonts. Initially computers could only handle English. This was then extended to other western-European languages and later to other major languages of the world. But there were no worldwide standards for thousands of minority languages and for some major languages written with complex scripts.

This lack of standards was true for:
- fonts that just did not have the needed characters,
- fonts that needed more than approximately 220 characters, and
- scripts (writing systems) that could not be conveniently represented by unchanging characters that were written from left to right.

If you were dealing with a script that behaved like Roman or Cyrillic scripts, the first of these problems was the easiest to solve. All you had to do was to change some of the characters that weren't needed. Unfortunately, because there were no standards, different font designers did this in different ways, producing incompatible fonts for similar groups of languages. SIL has been a major contributor to such nonstandard fonts, simply because SIL has been working in so many minority languages.

When fonts needed more than 220 characters, special approaches were developed for the three major Far-Eastern languages: Chinese, Japanese, and Korean. But there are many other languages that need more than 220 characters: Amharic of Ethiopia would be an example. The International Phonetic Alphabet falls into this category, even though it is not a separate language. Sometimes it is possible to build up characters so that the total number of characters would seem to be more than 220, but this is
only sleight of hand. SIL contributed to this through Tim Erickson's SDF system that was incorporated into Shoebox and LinguaLinks.

And then there remained the problem of complex scripts, such as scripts based on Devanagari or Arabic. The solutions were many and varied, and often tortuous. Again SIL contributed nonstandard solutions through the very flexible SDF system.

These customized approaches provided solutions, but they produced their own problems. I remember converting someone's Ph.D. dissertation from an MS-DOS document to a Windows document. There were four different fonts used under DOS, and each chapter had to use just one font. For one chapter, though, the chapter had to be printed once with one font and then the paper had to be reinserted for printing with a second font. This illustrates the ingenuity that was used to overcome the problems. But it was very inefficient. And in this instance, there was no easy way to archive the DOS document electronically to allow publishing years later.

Archiving is still a big headache. We need to store the many documents produced by our researchers over the years, but even with the advances that have been seen under Windows and on Mac computers, the fonts still have to be archived with each document—not a happy situation.

**Blessed are those who use Unicode!**

But a change is coming about. It is a change we cannot stop, and there are huge advantages associated with it. It is a change called Unicode.²

Unicode is a worldwide standard for all the writing systems of the world, including some ancient scripts. No longer will you confuse an eng (ŋ) with a Greek mu (µ) or a c-cedilla (ç) or a Cyrillic yu (ю) or an Arabic ghain (خ) or a Hebrew alef (א) or a Devanagari ka (क). I have not changed fonts when typing any of these characters;

Word 2000 found them all in the Times New Roman font, except, I think, the Devanagari ka. The Devanagari characters do not exist in my Times New Roman font, so Word went and grabbed the character from the Arial Unicode MS font that I have installed.³

Even the smiley face is from Times New Roman. To see all of the unusual characters above, readers will need Word 2000 with the Arial Unicode MS font installed from the Word 2000 CD. And they might additionally need Windows XP.

The old customized fonts⁴ only allowed approximately 220 different characters. Unicode was initially designed to provide some 65,000 characters, but this was later found to be inadequate, so the number has been extended to more than a million. Most of these are still empty slots, waiting to be assigned their characters.

Unfortunately, Unicode does not yet cover all the languages of the world, because the Unicode Consortium does not know about all the languages of the world. But they are getting there with help from people like you and me. At the present time, Unicode probably covers all the characters needed for 95 percent of the languages of the world, perhaps more, and, for these languages, each character cannot be confused with any other character. The character you see, such as the upper-case eng (Ŋ),

---

² For more detailed information about Unicode, visit the Unicode website at http://www.unicode.org/. See especially the comprehensive FAQ (Frequently Asked Questions) at http://www.unicode.org/unicode/faq/.
³ This might only fully work on Windows XP.
⁴ Sometimes known, perhaps incorrectly, as 8-bit fonts.
may not be the shape you would prefer, but this is a typeface issue—a different Unicode-compatible font may give you the shape you want.

In the past it was easily possible to receive a document but not have the correct font, so it was unreadable. But now for the vast majority of languages, if you use a modern operating system, Unicode-compatible software such as Word 2000 or later, and choose to use Unicode encoding for your document, any appropriate Unicode font will display everything correctly.

Until now, if you wanted to save your phonology write-up for posterity, you had to make sure that the reader or archivist had the correct fonts. But if your phonology write-up were a Unicode document, you should not need to provide any fonts. The reader would be able to choose from a number of appropriate Unicode-compatible fonts and see everything correctly.

You will even be able to display language materials on the Internet, showing the full orthographic details.

**What about the problems?**

Unicode sounds wonderful, but there must be problems. Sadly, there are. There is no such thing as a free lunch!

For most SIL teams, their old files are not in Unicode, and many files will need to be converted. This year (2002), data is being gathered from around the SIL world so that we can know what needs to be done. In each entity a computing specialist will be trained to provide the help needed for converting to Unicode.

Of course, there is no point in converting your data until there is software that can handle the scripts you need. Microsoft Word 2000 and later versions use Unicode and can handle many complex scripts, but perhaps will never handle all complex scripts.

Until now none of our specialist language programs could handle Unicode-compatible fonts and complex scripts, but that is expected to change this year with the release of several programs. From SIL expect WorldPad, a basic word processor, and the FieldWorks Research Data Notebook, designed mainly for anthropological research. From UBS expect Paratext; it should be available either late in 2002 or early in 2003.

Later releases of SIL's FieldWorks software will probably only allow Unicode data and will handle most complex scripts.

But what do you do if some of the characters you need are not in the Unicode standard? There are several possibilities here. First, you may find an alternative Unicode character that is acceptable to the local people. Second, a formal request can be submitted to the Unicode Consortium to have the characters added to the standard; but it may take several years before this becomes a useful solution for you. Third, there is a part of the Unicode standard that allows users to add extra characters to a private use area (PUA), but if you have to go this route, you will have to provide the font and send the font around with your data. Fourth, for complex scripts, the only option may be to customize a Unicode-compatible font, so breaking its compatibility. This should only be regarded as a temporary solution until the requisite characters have been added to the Unicode standard.

---

5 The FieldWorks Research Data Notebook will definitely handle Unicode, but might not handle complex scripts.

6 Through your entity Unicode representative, contact Peter Constable at the Non-Roman Script Initiative.
And what do you do if a script you need is not part of the Unicode standard? The Tifinagh script of Saharan Africa is one such. Well, it needs to become part of the Unicode standard, but it has sometimes been difficult for the Unicode Consortium to gather the information needed. If you need to use such a script, someone needs to put in a formal request to the Unicode Consortium.

When writing English, people use different typefaces to get different effects. I might choose the Times New Roman typeface, or the Arial typeface, or the Comic Sans typeface. No problem! I can choose whatever typeface I like, whatever suits my mood, and most English-speaking people will have no problem with it. But for some languages, the typefaces make a lot of difference. There are some languages that use Roman script, but require, for example, that accents must be sloping steeply; a fairly flat accent is not acceptable. For languages that use Arabic script, some typefaces are acceptable and others are not.

When it comes to keyboarding, you may ask how you can reasonably keyboard so many thousands of characters, especially since your keyboard may have less than 90 keys. Well, you will use a keyboard manager, such as Keyman, which has been rewritten to support Unicode. With it I have been able to create keyboards for complex scripts. You may be using the same font for all scripts, e.g., Times New Roman, but you may have keyboards for French, IPA, Hebrew, Greek, and Arabic. You will just switch to the keyboard you want and you will get the correct characters, and, as you type, the letters will be progressing in the correct direction and, if appropriate, changing shape as you go. It is almost magic!

Of course, a decision has to be made about the best time to change over to Unicode. It depends on a whole range of things: when the software you need is available; when the computer department in your entity is ready to facilitate the change; when the strategic partners you work with are ready to change; when someone has designed the keyboards you need; when there is an opportunity to get training; and so on.

You may be so close to the end of your project that there is no point in changing over to Unicode now. But you will eventually have to change your data to Unicode, to make sure that it can be adequately archived, and so be available for future revision, and for use by the language community and the academic world.

There is one other thing you need to know, another change that is coming, and it has to do with file formats. The two common file formats in the past have been Standard Format (SFM) files and Word files. There is a new worldwide standard called XML (eXtensible Markup Language). In the long term this is expected to completely replace SFM. XML will also be used extensively on the Internet. But XML can be the subject of another article at a more appropriate time.

Conclusion

Unicode is coming! You could resist it or ignore it, but it would be to your disadvantage to do so. For some translators, particularly those working with complex scripts, it will present possibilities that were previously only dreamt about. A more complete form of the International Phonetic Alphabet will be available. Anyone that uses a Roman script and requires combinations of diacritics will finally

---

7 A typeface may come in multiple font styles, e.g., regular, bold, italic, and bold-italic; one typeface, multiple font files.

8 For example, Windows XP (but choose Windows XP Pro for other reasons).

9 SIL’s Non-Roman Script Initiative (NRSI) is spearheading SIL’s transition to Unicode, and three NRSI staff members are official delegates to the Unicode Consortium. In addition, every SIL entity will soon have identified a Unicode representative to facilitate the entity’s transition to Unicode. If you have questions, or if you are ready to get started, then get in touch with your Unicode representative.
be able to produce documents that can be exchanged with others without being concerned about font
problems. Archivists will have a way of preserving data for future generations, and you will be able to
display your language data in its full orthographic detail on the Internet.

As an organization we have a new vision. For that vision to become a possibility we need
Unicode.