When to Convert to Unicode

Albert Bickford, Jim Brase, Lorna A. Priest,
SIL Non-Roman Script Initiative (NRSI)
2007-05-11

Linguists and others who work with minority language data often still use older special character systems (fonts and keyboards) that work only for certain languages. These older ('legacy') systems are gradually being replaced by Unicode (What is Unicode?[1]), which attempts to cover all languages in one unified system.

So, you are probably wondering: When is the right time to convert to Unicode? That's what this article is about. The question actually has two parts:

- When is it possible to convert?
- When is it necessary to convert?

There is usually a period of time—a window of opportunity—between the time when it becomes possible and when it becomes necessary. Ideally, you should plan ahead so you can convert your data during that window. You want to find a time when it is convenient to do so, before a conversion is forced on you at perhaps a very inconvenient time. Hopefully your window of opportunity will be large: it will be reasonable for you to convert your data to Unicode at least a year before it becomes necessary.

This article will help you recognize whether you and your language data have entered that window of opportunity. It will also help you predict whether you are approaching the end of the window: the point where a conversion is necessary. It will help you decide whether *now* is the time to convert.

Likewise, if you are just starting to work with a language, you may need to decide which special character system to use. From now on, most people should plan to start with Unicode, but there may be a few situations where older fonts and special character systems are necessary.

This article is written mostly for a general audience, although some of the links get more detailed and technical. At some point, you may need the advice of someone with special technical knowledge to decide when to convert, and most people need help to actually do the conversion and get set-up to use Unicode. This article will also help you decide when you need to get that help, and provide guidance to the technicians who are advising you. Because it is written for two audiences—ordinary users and technicians—the words "you" and "your" sometimes refer to the users/owners of the data and sometimes to the technicians assisting them. You decide whether they apply to you personally.

We start by listing the general questions you need to consider. You will need to click on the links to get to the sections that explore these questions in more detail.

Basic questions to consider

In order to answer the two basic questions above and define your window of opportunity, you need to ask yourself several more specific questions. These are listed in the following chart, with links to later sections that discuss them in more detail.

When are you ready for Unicode?	When is Unicode going to be forced on you?
If you can answer "yes" to *all* of the questions in this column, then you have entered the window of opportunity—you are ready to convert your data.	If you answer "yes" to *any* of the questions in this column, you have reached the end of the window of opportunity—Unicode has become a necessity for you.
• Is Unicode ready for you and your data? Does it contain all the special characters you need? (see section Is Unicode ready for you?[2])	• Are you working under the authority of an organization that requires you to use Unicode? (see section Is anyone requiring you to use Unicode?[7])
• Are there fonts that contain the characters you need, and are suitable in all other respects? (see section Are there	

[1] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ1

[2] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ2

fonts available that will work for you?[3]) • Are there input methods (e.g. keyboards) that make it easy to type the characters you need? (see section Can you type all the characters you need?[4]) • Is the technical expertise necessary to do the conversion available to you? (see section Is the technical expertise to do the conversion available to you?[5]) • Are you and others who work with your data ready to learn about how to use Unicode? (see section Are you ready to learn about how to use Unicode?[6])	
• Consider the different things that you want to do with your linguistic data; will your operating system and your application software allow you to do those tasks if you use Unicode? Or, to put it another way, does available software that handles Unicode do everything that you need it to? (see section Does available Unicode software meet your needs?[8])	• Consider all the software that you need to use: Does any of it require Unicode? (see section Is software going to force Unicode on you?[9])
• Do you have *no* need to exchange data with other people who are using your 'legacy' character set? (see section Do your colleagues use Unicode?[10])	• Do you need to exchange data with other people who use Unicode? (see section Do your colleagues use Unicode?[11]) • In particular, do you need to archive your data as Unicode for long-term preservation? (see section Is it time to archive your data?[12])

• Are you willing and able to "straddle the fence"—working with older systems and Unicode at the same time? (see section Are you willing and able to "straddle the fence"?[13])

Basic questions to consider

This table may provide enough detail for some people to decide whether they are ready to convert to Unicode. If so, you can stop here and get on with converting to Unicode. But, if you need more details, read the linked pages.

---

[3] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ3

[4] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ4

[5] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ9

[6] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ10

[7] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ6

[8] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ5

[9] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ7

[10] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ11

[11] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ11

[12] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ8

[13] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ12

With the exception of the Table of Contents, most links in the document will take you to the website, rather than to a point in the document.

What is Unicode? and Why do I need to use Unicode?

Unicode is a character encoding standard that has widespread acceptance. Microsoft uses Unicode at its core. Whether you realize it or not, you are using Unicode already! Basically, "computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters.[14]" This has been the problem we, in SIL, have often run into. If you are using a legacy encoding your font conflicts with the font someone in another area of the world uses. You might have an ɬ in your font while someplace else someone used a Σ at the same codepoint. Your files are incompatible. Unicode provides a unique number for every character and so you do not have this problem if you use Unicode. If your document calls for U+0289 ʉ it will be clear to any computer program what the character should be.

Advanced Resources

- An Introduction to Unicode[15]
- Understanding Unicode[16]

Is Unicode ready for you?

Unicode contains a huge inventory of characters, currently over 100,000. But, that doesn't guarantee that it will have every character that you need. Now, there has been a major effort over almost two decades to identify all the characters that need to be in Unicode, and most people will find that their data can be represented in Unicode with no problem. But, if the language you work in happens to have one of those rare characters that hasn't yet been added to Unicode, or worse, a whole script that isn't yet included, then the first order of business is to request that it be added.

What do you do if there is no established orthography for the language? That, of course, gives you much greater flexibility. As long as you choose from among the characters already available in Unicode and follow standard conventions in how those characters are used, there should be no problem. For more details on this subject, see Orthography development in relation to Unicode[17].

But, most people reading this article need to work with an established orthography. So you need to check to see if Unicode will support it. Most major languages are already fully-supported, but minority languages may not be. So, make a list of all the characters that you need to use. You will probably need help from a Unicode expert to know if the characters are in Unicode, but you can get started on your own by listing out what you need, then take your list to the expert.

Here are some of the issues you should consider:

- Upper and lower case
- Diacritics
- Borrowed words
- Punctuation and other symbols

[14] http://www.unicode.org/standard/WhatIsUnicode.html

[15] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_6_2

[16] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-Chapter04a

[17] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=OrthographyDev

- Phonetic transcription
- Other languages and scripts you may use

More advanced:

Inventorying the Character Set and Comparing it to Unicode:

- Include upper-case as well as lower case. Some people think that if a character never occurs first in a word, they don't need an upper-case version of it. Then, they run into a situation when they need to use all-caps. If your writing system regularly makes a distinction between upper and lower case (or any analogous difference in a character's appearance that is not predictable from context), list an upper-case version of it.
- On the other hand, if the shape of the character changes based on its immediate context, then also list all the variant shapes. For example, in Arabic-based scripts, letters change shape depending on whether they occur first, middle, or last in a word. Unicode handles this by treating all the variant shapes as the same character and relying on smart fonts to give the right shape in each context. (Click here[18] for more information.)
- Include diacritics in your list. List all possible combinations of diacritics with base characters (again, remember to include upper-case) as well as combinations of two or more diacritics on the same letter. Unicode does provide ways to form arbitrary combinations of base characters and diacritics, but the most common combinations are also available pre-assembled as single characters. So, you'll want to check if these "pre-composed" characters are available for the combinations that you use. If there aren't, then you need to make sure that the diacritic is included as a separate character that can be combined with other characters.
- Include any characters that only occur in borrowed words.
- Include punctuation characters and any other special symbols other than ordinary word-building characters. Almost certainly, they will be included, but if you use anything unusual, that isn't in a major language, you should check this out carefully.
- Consider all languages that you work with, including those that you may only use occasionally. Major languages are already included, so that's not a problem. But, if you want to exchange data with people working in related languages, you may need characters for those other languages.
- Consider symbols you need for phonetic transcription. All symbols currently approved by the International Phonetic Association (IPA)[19] are already included. However, if you use a different transcription system, such as Americanist phonetic characters or phonetic symbols that are only used in a particular part of the world or a certain language family, you need to check.
- If any language that you work with has more than one script, consider each script separately.
- Don't plan to depend on formatting such as underlining, superscripting, or italics in order to represent your characters. For example, if you need to represent a superscript ^{h}, don't plan to use an ordinary h and just apply superscripting to it. Besides being clumsy to type, this method is unreliable. If all the formatting ever gets stripped off your text, then the distinction between h and ^{h} will disappear. You will

[18] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ3

need to represent these two as separate characters in Unicode. (And, yes, Unicode does have a separate character for ɦ.)

- You don't need to worry at this point about fine details of appearance, as long as the character in Unicode is recognizable the same character as the one you are using. For example, some languages prefer an upper-case eng (Ŋ) that is just a larger version of the lower case eng ŋ; others prefer one that looks like a regular upper case N with a tail (Ɲ). These two shapes are considered "glyph variants" of the same character in Unicode and are represented the same way. You control which version you use through the fonts that you use.

- You *do*, however, have to pay attention to how a character is used. Unicode sometimes includes more than one character with the same appearance. For example, there is an apostrophe that is used for punctuation and a separate apostrophe that is used to represent glottalization. The two characters look alike, but one is a punctuation mark and the other is a word-building character. You might have thought of them as the "same" character until now, but they function differently in a writing system, and makes a difference for functions like selecting whole words, breaking lines, searching and sorting. So, you can't just pick a character out of a Unicode chart because it looks right; you have to read the descriptions of each character to make sure you've got the right one. If you've been representing two characters the same way up until now, then in the process of conversion to Unicode, you will need to figure out some way to distinguish them. Besides the difference between punctuation marks and word-building characters, you also need to distinguish characters that are used as diacritics vs. ones with the same appearance which are full characters on their own.

- In general, Unicode itself is only concerned with the computer being able to recognize a character reliably. Matters of appearance, such as variant shapes of letters in different contexts, preferences about letter shapes such as "a" vs. "ɑ", and fine positioning of diacritics are not distinguished in Unicode itself, because they are either predictable from context (and thus should be handled by smart fonts), or because they are a matter of personal preference (and thus should be handled as formatting, i.e., by choosing what font you want to use to display the data or choosing options within the font).

In listing out the characters you need, you may find it helpful to look at the inventory of characters in the fonts that you are currently using. In general, you should verify that all of them are in Unicode. (If you are using some ISO standard character set, like the standard Windows Latin fonts, or Big 5 for Chinese, then all those characters are in Unicode.)

Now, it could be there are characters in an old custom font that you never use. As long as you can guarantee that they don't ever occur in your existing data, you don't need to worry about them being in Unicode. But, if there is any doubt (for example, if they might have been typed by mistake), then it is best to plan to convert them to Unicode along with everything else.

After doing this inventory and consulting with a Unicode expert to make sure all the characters you need are in Unicode, what happens if some characters are missing? If so, you have three options:

- Decide to discontinue using the characters that are missing and use something else that *is* in Unicode. You can't always do this, of course, but sometimes it is the best option.

- There may be a font available that includes the character you need in its "private use area[20]", a section of Unicode that is intended for local customization. Or, you may be able to arrange to have the characters you need added to the private use area[21] of a font.

[19] http://www.arts.gla.ac.uk/IPA

[20] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=glossary#pua

- Get the missing character(s) or scripts into Unicode.

The last two options require consulting with a font designer and/or someone who is in regular contact with the Unicode consortium. If you are in SIL you may contact the NRSI for help in making Unicode proposals (go to Insite, find the NRSI InfoCenter, click on the Resources tab and look for "NRSI Resource Personnel"). Non-SIL may find help through the Script Encoding Initiative[22].

Advanced Resources

- Adding new characters and scripts to Unicode[23]
- Orthography development in relation to Unicode[24]
- Guidelines for Writing System Support: Roles and Actors[25]
- Creating a Chart of Your Legacy Mapping[26]

Are there fonts available that will work for you?

Just having the characters you need in Unicode is not enough. Unicode is too big to fit all in one font, so you also have to find out which fonts contain all the characters you need. That is, all the characters you need for a particular language and script should be in the *same* font, and hopefully you'll find more than one. If you use multiple languages or scripts, you can use a different font for each one.

There are some fonts that contain very large inventories of Unicode, plus many others that focus on specific scripts, so you've got a good chance of finding a font that will work. More are appearing every year. See Font Resources[27].

When you're looking for a font, however, you need to consider more than just whether the font has the characters you need. For example:

- Smart behavior, in which the appearance of letters changes dynamically depending on context.
- Variant shapes of letters, such as "a" vs. "ɑ".
- Overall appearance (serif vs. sans-serif vs. fixed-width, formal vs. informal, etc.)
- Licensing restrictions
- Special requirements for high-quality publication

Advanced Topic

Choosing a Font

- If you need any sort of smart behavior, such as correct positioning of diacritics, forming ligatures, changing the shape or position of a letter depending on its context, then you

[21] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=glossary#pua

[22] http://linguistics.berkeley.edu/sei/

[23] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_6_3

[24] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=OrthographyDev

[25] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_3

[26] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTTLegacyMap

have to make sure that the font will provide that behavior in combination with your operating system and application software as discussed later in section Does available Unicode software meet your needs?[28].

- If you have preferences about the shapes of letters (such as “a” vs. “ɑ”), make sure all the shapes you want are available in the same font. (Sometimes font designers provide more than one shape for the same letter in one font, but you need special software in order to access anything other than the default letter shape.)
- You may have other requirements for the overall appearance of the font. For example, you may need a fixed-width font or a font suitable for new readers.
- Fonts sometimes have licensing restrictions that don’t allow you to do everything you might want with them. For example, you might be able to print with them on your own computer, but not transfer them to anyone else for them to view, print, or edit the files. The best choice is fonts that are covered by a license that grants you a lot of freedom, such as the Open Font License[29].
- Not every font works for high-quality publication. Appearance may be adequate for day-to-day work, but not meet the standards of your publisher. In particular, you will need a font family that has all four standard styles (regular, italic, bold, and bold italic). Application software or operating systems sometimes simulate bold and italic styles even if bold and italic fonts are not available. But, for high-quality publication, all four fonts are needed. This need not prevent switching to Unicode if you aren’t going to be doing any high-quality publishing in the next few years, but you should find out what options are being developed to meet your needs in the future.
- Some universities and publishing houses require their authors to use particular legacy fonts that are not Unicode-compatible, such as SIL’s IPA93 fonts (SIL Encore IPA Fonts[30]). If you’re subject to such restrictions, you may have to either delay converting to Unicode or convert your data back to their specifications when you give it to them. (See section Are you willing and able to “straddle the fence”?[31].)

If you find that there are no fonts that meet your needs, you have these options:

- Abandon use of the feature that isn’t available. It may be more important to convert to Unicode than to retain some feature that is only supported by an older special character system.
- Convert to Unicode anyway and put up with the missing feature for a few years while you wait for fonts to support it.
- Work actively with font designers to implement the features you need.

If our fonts do not quite meet your needs *and* if you are in SIL you may contact the NRSI for help (go to Insite, find the NRSI InfoCenter, click on the Resources tab and look for "NRSI Resource Personnel").

Advanced Resources

- Font Resources[32]

[27] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=CatTypeDesignResources

[28] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ5

[29] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=OFL

[30] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=encore-ipa

[31] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ12

- Glyph Design[33]
- Rendering technologies overview[34]

Can you type all the characters you need?

Fonts are just one half of what is needed for supporting special characters. You also need to have some easy way to type them, either specialized keyboards that allow you to type all the characters you need or (with some languages) more elaborate input methods called Input Method Editors (IMEs). Here's where you'll probably need to consult a computer technician to find one that will give you the characters you need and will work with all the software you want to use (see section Does available Unicode software meet your needs?[35]). For more details about options and some of the issues, see Some tools and resources for character input[36].

You probably are used to keyboarding characters in a particular way. It may or may not be possible to continue typing things in exactly the same way after you convert to Unicode. For example, if you are used to typing diacritics before base characters as dead keys, you may need to learn to type them afterward. Learning isn't really very hard, but it does take some time and attention and the willingness to do so. If you're not prepared to learn a new keyboarding system, then find out if a new keyboard can be built for you that allows you to use the same keystrokes you're used to using.

Advanced Resources

- Keystrokes and codepoints[37]
- Data Entry and Editing[38]
- An introduction to keyboard design theory: What goes where?[39]
- Some tools and resources for character input[40]
- Comparing Keyman and Microsoft Windows Keyboard Layout Creator[41]

Does available Unicode software meet your needs?

For Unicode to work, the operating system and other programs you use need to know how to use it correctly. This includes word processors, dictionary and interlinearizing tools, specialized translation editors, etc. More and more software is becoming available every year that supports Unicode, and the whole software industry is moving in this direction. However, you may need to do things that are still only possible in software that requires using older fonts and special character systems. So, before you switch

---

[32] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=CatTypeDesignResources

[33] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_8

[34] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-Chapter07

[35] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ5

[36] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=inputtoollinks

[37] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_5_3

[38] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_7

[39] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=KeybrdDesign

[40] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=inputtoollinks

[41] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=KeymanVsMSKLC

to Unicode, you need to make sure that you'll still be able to do everything that you want to with your data afterward.

To do so, make a list of the different programs that you currently use (or want to start using). What, specifically, do you use those programs for? Do any of them have any special capabilities that you need? Consider especially the following categories of software:

- Word processing
- Graphic editing (if you need to include text in your graphics)
- Dictionaries and interlinear text
- Specialized editors, e.g. for translation
- Language survey, especially wordlist comparison
- Desktop publishing (consider the full range of materials you want to produce, including both technical publications for scholars and vernacular materials for native speakers, whether you write them from scratch or adapt a "shell book" from some other language)
- Output to printers, PDF, and web pages (HTML)
- High-end publishing (you probably won't do the work, you need to know that your publisher has the necessary software)
- Any other software that you use with minority language data.

In doing this inventory, you may find it helpful to use a spreadsheet like the one found here: Creating a Chart of Your Legacy Mapping[42].

Then, talk with someone who knows about these different categories of software to find out if there is Unicode-capable software that will do what you need. In deciding this, you need to consider the program itself, the operating system you are using, and the various options for fonts and keyboarding. Everything has to work together properly. Plus, some software supports Unicode only part-way; you'll need to decide if that's good enough. For details about different combinations of software and levels of Unicode support, see Software requirements for different levels of Unicode Support[43] and Applications that provide an adequate level of support for SIL Unicode Roman fonts[44].

If the particular software you're used to doesn't support Unicode, look for some similar tool that will allow you to do the same things with Unicode. You'll probably need to learn some new software when you convert to Unicode; sometimes new versions of familiar programs, and sometimes totally different programs. It takes some effort, but in the long run it will be worth it for most people.

Then, finally, you have to decide whether the newer Unicode software will work on your computer. You may need to upgrade to a new computer in order to run the newer software. Further, you need to think about all the computers you want to use. It is generally not a good idea to convert to Unicode on one computer but leave data in an older special character system on another computer—the potential of getting them mixed up is too great. (For more on this, see section Are you willing and able to "straddle the fence"?[45])

Advanced Resources

- Determine the needs[46]

---

[42] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTTLegacyMap

[43] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UnicodeSupport

[44] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=Complex_AdLvSup

[45] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTConvertQ12

[46] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_2

- Software requirements for different levels of Unicode Support[47]
- Applications that provide an adequate level of support for SIL Unicode Roman fonts[48]
- Creating a Chart of Your Legacy Mapping[49]

Is anyone requiring you to use Unicode?

So far, there are few organizations that actually require people to use Unicode. However, we can anticipate that as the use of Unicode becomes more widespread, it may become more common in the future for universities, scholarly societies, NGOs, or governments to require this of researchers or language development workers under their jurisdiction.

Advanced Resources

- Guidelines for Writing System Support: Roles and Actors[50]

Is software going to force Unicode on you?

Some of the newest software that is coming out will *only* support Unicode. At that point, you're forced into a hard choice: continue to use your old software with your old special character tools, or switch to Unicode to take advantage of the power of the newer programs.

Actually, to be precise, it may still be possible to fool some of the newer programs into working with older special character systems, but this is a risky business. At any point, a new version of the program may come out that will no longer be fooled. At that point, things will start going wrong, like characters showing up as little boxes, or lines breaking in the middle of words, or being changed to other characters when you copy-and-past them to other programs, or keyboards no longer working. Now that many programs update themselves automatically over the internet, things may break even when you aren't aware that anything has changed.

You may decide that "the old is good" and decide to keep on working with the same software as you currently use so you don't have to convert your data to Unicode. This is a reasonable option for a few years, but as time goes on it gets more and more risky. If your computer hardware breaks down and can't be fixed, you may have to upgrade to a new computer that won't run your old programs. As time goes on, there will be fewer and fewer people who know the old software and can help you with it. Your data may even become stranded in an old file format so that you can't print it or do anything to it. So, at some point you will probably be forced to change. If you are almost done with your work on a language, you may manage to avoid changing—someone else can convert your data to Unicode for archiving and help other people work with your data using newer software. But, then again, these sorts of problems have a way of creeping up and ambushing us at awkward times. Don't just ignore that possibility and pretend it can't happen to you. Make a conscious choice about accepting the risks involved and have a backup plan for converting to Unicode in a hurry if that becomes necessary.

Advanced Resources

- Software requirements for different levels of Unicode Support[51]

---

[47] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UnicodeSupport

[48] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=Complex_AdLvSup

[49] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTTLegacyMap

[50] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=WSI_Guidelines_Sec_3

[51] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UnicodeSupport

Is it time to archive your data?

Over the long run, Unicode is going to become the only system for special characters in the world. In ten or twenty years, it is unlikely that any other data will be readable. So, when it comes time to archive data for people to use in the future, Unicode is really the only reasonable option.

That doesn't necessarily mean that you need to convert the data to Unicode yourself. Archives are generally happy to accept data in any form, as long as it meets certain minimal standards. So, if you are using some non-Unicode system, it can be contributed to the archive in that form, as long as there is enough information about your system that someone in the future could convert it to Unicode. Usually this means including a copy of any custom fonts you use plus a file that describes what characters are in the font and how they are coded.

Still, there are real advantages to converting the data now. After you are gone, there may be no one who knows the language well enough to check the result of the conversion to make sure nothing got garbled in the process. Or, there is the danger that data will sit in the archive unconverted for so long that today's conversion tools no longer work or there is no one with the expertise to convert it. So, if you have the time and energy to work through the process, and there is a technician available to help you, it is better to do it now.

Is the technical expertise to do the conversion available to you?

Actually converting data to Unicode requires special expertise. Some users can learn what is required. There are a lot of tools and instructions available, such as the following:

- How to Write a Conversion Mapping for your Legacy Font[52]
- SILConverters 2.5[53]
- Encoding Conversion Frequently Asked Questions[54]
- SIL IPA93 Data Conversion[55]

But, most people will want to work with someone who is specially trained in the process. Or, if you want to do it yourself, at least get advice from someone with experience, and maybe have them check your work, so as to avoid the biggest pitfalls.

Here's what's involved in doing a conversion, once you've made the decision to go ahead with it:

- Locating all data that needs to be converted.
- Deciding what techniques and tools are appropriate for each type of file.
- Customizing those tools for your particular situation. This usually includes obtaining or writing a "mapping table" that indicates how each of the characters in your old system should be transformed into Unicode. Special procedures may be needed if more than one special character system is used in the same file, or the conversion rules for borrowed words are different from native words, or files are in unusual formats.
- Actually doing the conversion.

---

[52] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTTWriteMap

[53] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=EncCnvtrs

[54] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=EncConvFAQ

[55] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=SILIPA93DataConversion

- Checking the result visually to make sure nothing got garbled in the process, and redoing the process if necessary.
- Archiving all files that are in the older special character system and removing them from any place where they may be confused with the Unicode files.
- Obtaining new Unicode fonts and keyboards and installing them on all computers that need them.
- Obtaining new Unicode-capable software and installing it.
- Learning to use all the new software.

Advanced Resources

If you feel ready for more information in the area of encoding, you may want to read:

- Understanding characters, keystrokes, codepoints and glyphs[56]
- Character set encoding basics[57]
- Mapping codepoints to Unicode encoding forms[58]
- A review of characters with compatibility decompositions[59]
- Character Encoding Choices in Paratext 6[60]
- How to Write a Conversion Mapping for your Legacy Font[61]
- SILConverters 2.5[62]
- Encoding Conversion Frequently Asked Questions[63]
- SIL IPA93 Data Conversion[64]

Are you ready to learn about how to use Unicode?

Sometimes, a conversion to Unicode is almost completely transparent to a user. They'll be able to continue working almost the same way as they always have. However, it doesn't always work that way. As you think through when to do a conversion, find out how much you'll have to learn. This may include

- Which fonts and keyboard to use with Unicode data.
- How to type the characters you need.

---

[56] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-Chapter02

[57] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-Chapter03

[58] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-AppendixA

[59] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-AppendixB

[60] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=CharEncInPT6

[61] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=UTTWriteMap

[62] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=EncCnvtrs

[63] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=EncConvFAQ

[64] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=SILIPA93DataConversion

- If you need to code diacritics as separate characters, how to work with them in various programs.
- How to distinguish two characters that look the same but behave differently (and how to train yourself to type them correctly).
- How to recognize when something is wrong and what to do or who to call to fix it.
- How to use new Unicode-capable software.

Most everyone can learn these things. It doesn't take a lot of time, but it will slow you down for a couple weeks or in rare cases a few months. So, don't do a Unicode conversion when you have pressing deadlines coming up, if you can help it. Plan ahead for a time when the conditions are right for an unhurried change-over.

Do your colleagues use Unicode?

This question is unlike the others, because it relates to both ends of the window of opportunity. If you need to exchange data with people who have not made the switch to Unicode, then that may be reason for you to wait until they are ready to convert at the same time as you are. On the other hand, if you need to exchange data with people who have already made the switch, then it may well be time for you to join them. The question is especially important if you have to share data back and forth frequently, especially with people who are working together with you on the same project.

Now, we should clarify what we mean by "exchanging data". If the other person just needs to be able to read your data, then almost always there are ways to make that possible. Most computers and operating systems support Unicode to some extent, so your Unicode data should be readable on other computers. Or, if you have legacy data and want to share it with people who use Unicode exclusively, you'll have to remember to provide them with your fonts. (If nothing else, you can package the data in a PDF file with embedded fonts; that almost always works under both scenarios.)

But, if people need to merge your data with theirs, or you both need to make changes to the same files, then you are going to have to use the same special character system. Either that, or you will need a way to reliably convert the data back and forth, but that is the subject for the next section.

A special case of this question concerns publishers. If you are going to publish an article, a dictionary, or other item involving linguistic data, your publisher may have requirements that are different from yours; they may not be ready to use Unicode or they may require it. In that case, you will either have to convert your data to match their requirements, or they will, or you will need to find another publisher.

Are you willing and able to "straddle the fence"?

This question also relates to both ends of the window of opportunity. If you have the technical know-how or technical support to work with more than one special character system at the same time, then the window of opportunity expands considerably. You have the option to start using Unicode on some of your data sooner, and you can postpone the point when you must leave legacy encodings behind. So, if you can answer "yes" to this question, you will have more flexibility as to when to convert. Unfortunately, however, not everyone can answer "yes", and there are enough potential problems to this approach that the majority of users and situations should probably answer "no".

Here are some possible scenarios:

- You use Unicode on one computer and the older system on another computer. Either you don't move data back and forth between the computers, or you learn how to convert the data back and forth as needed.
- Some software allows you to store the data on the hard disk in the older system, but will automatically convert it to Unicode when it reads it into the program, then converts it back to the older system when you save the file.

- You use different special character systems in different files on the same computer, and convert back and forth as needed.

All of these can be problematic. It can be easy to mix the two systems up. For example, you might start working on a file using the wrong keyboard and font, and introduce a mixture of Unicode and non-Unicode data in the same file. You may want to copy-and-paste data from one file to another, but this won't work if one uses Unicode and the other doesn't. If you have to convert data to and from Unicode, you have to learn how to use the conversion tools and you can spend a lot of time converting data instead of getting real work done.

Still, for some users who understand the issues and can use the necessary software, it can be convenient to convert to Unicode in stages over a period of time, rather than all at once. You'll have to decide if you are one of those, and whether the advantages of maintaining two systems outweigh the costs.