Unicode Strategy:

Syllabics and Computing in the 21st Century

Bill Jancewicz, October, 2001
SIL / NDC

Introduction

What's Unicode?

In order to address the interchange of texts across platforms, as well as to provide the number of characters required by the languages of the world, an industry standard character set encoding named Unicode was developed. The Unicode standard is able to support over one million characters, and has been designed to support all characters from all scripts worldwide.

Both Windows and Macintosh operating systems have taken steps in the late 1990s toward implementing Unicode, and at the time of this writing (2001) that implementation provides substantial support for both platforms. That is, Macintosh OS X and Windows 2000 both provide adequate support for the Unicode standard. While there are as yet only a few applications supporting Unicode on these platforms, it is evident that the IT (Information Technology) industry is moving toward compliance with the Unicode standard in the very near future.

This shift toward Unicode will have an impact on everyone who uses computers and requires characters or symbols outside the standard character sets, including phonetic characters and mathematical symbols. Obviously, users of Cree and Naskapi Syllabics will be affected.

Unicode and Canadian Syllabics

Over the years, various approaches to addressing the problem of using Non-Roman scripts have been implemented. These various approaches were all limited by the available hardware and software in the communities where these scripts were needed. The Naskapi community has benefited by having a linguist who was also familiar with some script technology for computers living in the community to develop the solutions required. These solutions were consistently upgraded to suit the available technology.

For example, from 1989 to 1993, the available technology in the Naskapi community included computer platforms that supported the MS-DOS operating system. Extensive collections of bit-mapped fonts were developed to support the various printers and computer screens that were in use in the community. This was the first local application of computer technology to Naskapi. Among other publications, the Naskapi Lexicon was produced using this technology.

In 1994, Windows 3.1 began to be used in the community, and TrueType Naskapi fonts were developed to support this platform. A custom encoding was designed, and then revised, until the local "standard" encoding "Naskapi ANSI" was established in 1995. These fonts provided much higher quality publications in the Naskapi language, with the ability to scale and format Naskapi texts.

Two other encodings were also developed for the Macintosh computer, but these were necessarily different from the Windows encoding because of the incompatibility of the operating system.

From 1995 until the present, these three different encodings have been used to type Naskapi Syllabics on Windows and Macintosh computers. This meant that normally, any material keyboarded on one kind of computer could not be used on the other kind of computer without extensive and often time-consuming data conversion.

For example, if a publisher or translation client used specialized Macintosh publishing equipment, while the Naskapi translators at the Nation office or the NDC use Word for Windows for keyboarding, the files could not be used readily by the client without this data conversion process. However, for day to day "in house" translation and printing of Naskapi documents, each user using his or her own system found their encoding adequate for the job. For this reason, the Naskapi school operated the Curriculum Development project using Macintosh computers, while the rest of the community uses Windows PCs.

To further complicate matters, other First Nations communities, including the James Bay Cree of Quebec also use syllabics but do not use an identical orthography to the Naskapi. Further, they relied on their own computer specialists and commercial facilities to develop their own solutions to the syllabics challenge, again out of necessity using a different encoding for Windows and Macintosh. This scenario ran its course in dozens of Native communities across Canada, and indeed in minority language groups around the world.

It was to address this "Babel" of multi-lingual computer encodings that Unicode was developed.

In Canada, representatives of the various language groups affected had already been meeting in order to address the computer encoding challenge on their own, and launched the Computer Coding for Aboriginal Language Syllabics (CCALS) project under the Federal Department of Communications, in order to provide community-level guidance for the Canadian Standard Association (CSA) and the International Standards Organization (ISO) in Geneva. By 1992, the Canadian Aboriginal Syllabics Encoding Committee (CASEC) was formed, and in its 1994 report provided the CSA and ISO with the basic syllabic repertoire which ultimately became a part of the Unicode Standard. Cree language techicians from Cree Programs participated in CASEC, as did representatives from the Naskapi community.

So, while the Unicode Standard has developed and provided a standard encoding for all the syllabic character sets in Canada, the actual hardware and software in use in the communities still (as of 2001) uses the non-standard "local" solutions, referred to as "Legacy" encodings.

This is also the current Cree and Naskapi situation.

Information Technology in the 21st Century

Desktop Computers in the 1990s

We have seen remarkable growth in the Information Technology (IT) business in the 1990s. From computers with only 640K bytes of memory and 30 Mb Hard disks with speeds of 8MHz, we now commonly use computers with more than 128 Mb of memory, 30 Gb Hard disks, and speeds above 800 MHz. Also, computers are connected by networks across the room and the Internet across the world. This increase in computing power and capacity also permits and requires far more sophisticated software.

Software and Operating systems in the 1990s

The Windows and Macintosh graphical user interfaces (GUI) have helped to make sophisticated computer applications accessible to a wide range of users. They have also allowed incredible flexibility in the range of characters, shapes and colours that may be displayed. This flexibility

has permitted extensive multi-language utility: languages with any alphabet can be implemented on computers, with high-quality publishing capabilities. However, adequate support for a unified character encoding scheme was not implemented by the operating systems and software until the year 2000.

Implementation of the Unicode Standard

With the introduction of Windows 2000 and the Macintosh operating system OS X, the IT industry finally has system-level support for Unicode. For Canadian Syllabics, this means that the orthographies that formerly had to be implemented using "hacked" font and input (keyboard) Legacy encodings, could now be (for the most part) seamlessly integrated into the computer system.

However, while it is essential that the underlying operating system include support for Unicode, it is also necessary for the individual applications to support Unicode as well. For example, even though the computer is running Windows 2000, which supports Unicode, if the user is still using a word processor that does not, like Microsoft Word version 6.0, then Unicode encoding does not work.

The IT industry is just beginning to support Unicode in applications. All versions of Microsoft Office for Windows has supported Unicode since Office 97. But there are still applications in use that will not support Unicode.

Computer users with minority-language requirements, like the Canadian Syllabics users of Cree and Naskapi need to have a workable plan or strategy to prepare for the shift to Unicode that has already begun.

Strategy for Unicode Transition

Stage 0

Stage 0 is the "status quo", that is, however Canadian Syllabics are implemented that currently "works", using Legacy encodings. The computer users need to assess and become aware of their own system, how it works, how it is different from neighboring languages using syllabics, and how it is different from Unicode. This assessment is crucial for a smooth transition.

It means keeping a character encoding inventory, as well as an accounting of the kinds of documents and software applications that are being used to process syllabic characters.

Stage 1

This stage begins to find solutions. It begins when users have a need to transfer data from one operating system to another, or need to move to newer applications and systems that no longer support the "hacked" Legacy font solutions, or a need to put syllabics on the World Wide Web.

In order to meet these needs, a workable "round-trip" conversion process must be developed. That is, some kind of computer program must be written that permits documents that were produced using the older Legacy fonts to be converted to Unicode encoding, without any loss of data.

It is also necessary to maintain computer systems active that can do ***both*** Unicode and the Legacy fonts. Windows 2000 is one such computer system. It is positioned to be the transitional system on which these conversions take place. Windows 2000 can still adequately support most Legacy font solutions. That is, everything that already "works" can still work on Windows 2000.

Windows 2000 also supports the new paradigm of Unicode.

The fonts and keyboarding programs for ***both*** the Legacy and Unicode encodings should be installed, and users should be trained in their use and distinctives, as well as converting back and forth between them.

The program to convert between the two encodings should also be installed and documents for each encoding kept in separate folders. For Cree and Naskapi, the conversion program can be a toolbar button in Microsoft Word that calls a Visual Basic routine that will perform the change to Unicode and back. Thus, users can open a document, check the encoding, and make the change if necessary. I have already developed such programs for Naskapi and Cree.

Stage 2

This stage is the systematic implementation of the solutions developed in Stage 1. Every document of archival quality should be converted to Unicode encoding during this stage. Backward compatibility should be tracked: that is, if the need arises for a document to be returned to the older Legacy encoding, the procedures to carry this out should be maintained. Some applications will still require the older Legacy encodings. Most new applications can handle documents in Unicode, so any documents created with these applications can be done in Unicode without any conversion necessary.

Stage 3

This stage occurs as the older systems and applications that require the Legacy encodings get phased out. When a Unicode-aware program is installed that replaces an older non-Unicode program, all the documents used by that application can be converted to Unicode and then the non-Unicode version of the software can be removed from the system. This stage continues until there is no longer any need for the older Legacy encodings.

Stage 4

Once all the documents needed by the user have been converted to Unicode, and there is no longer any need for the non-Unicode applications, the transition will be complete.

Snapshot: 2001

In 2001, ***Unicode*** is necessary to send and receive e-mail in syllabics, to do all interactive Internet and World Wide Web applications in syllabics.

But it is possible to continue to use word processing and desktop publishing applications for syllabics in *either* Unicode or Legacy fonts.

Further, there are still important software applications used for syllabics that do NOT yet support Unicode, like Shoebox (for Native language dictionaries) and Paratext (for Native language translation). Given the scenario in 2001, we can see that we are still in “Stage 1” to “Stage 2”. Some Unicode font support has been developed, although not yet as extensive in terms of typeface variety as the Legacy fonts, and some conversion programs between Unicode and Legacy encodings have been written. To do Web and e-mail applications, Unicode is required. To do dictionary and translation applications, the Legacy fonts are required. Word processing and publishing straddles the fence; it can be done in either Unicode or Legacy.

What to do Next

Communication

Users and decision-makers involved in Non-Roman scripts on computers need to be aware of the current situation, and know where they are in the transition strategy. Users need to know how to make conversions from one encoding to the other, and know how to tell them apart. Decision-

makers involved in purchase and deployment of equipment and software need to be aware of the special needs required by this transition, and make their decisions accordingly.

For example, while it is true that Macintosh OS X is the only Mac operating system that has full Unicode support, currently there are very few Unicode aware applications available for this system. A decision-maker should be aware of this, and not specify Macintosh computers to handle syllabics at this stage of the transition. Macintosh computers are currently workable for "Stage 0", and may be well suited eventually for "Stage 3" or "Stage 4". But to make a transition, only PCs running Windows 2000 or XP (and a recent version of MS Office: 97, 2000 or XP) should be the considered.

Training

Computer users will be facing these changes soon, if they have not already. There is no substitute for adequate training and practice in the techniques of performing the conversions between encodings and in recognizing which encoding is appropriate to each application.

Consultation and Collaboration

The reason Unicode exists is to unify and standardize the multi-lingual computing world. There are lots of other people and organizations facing the same thing that the Cree and Naskapi communities are facing with their various "hacked" font solutions, Legacy applications, and the growing need to share their documents across platforms or across the Internet. Consultation with experts in this field is an important facet of this journey, and can really help a lot. SIL International has the Non-Roman Script Initiative, which provides access to much of this expertise. Also, linguistic and general computing help is an absolute necessity. By working together, we can begin to benefit early from the integration of Unicode to our Native Language computing needs.

Contact Information

You can contact Bill Jancewicz using any the following means:

Mail: Bill Jancewicz
Naskapi Language Studies / SIL
Naskapi Development Corporation
P.O. Box 5123
Kawawachikamach, Quebec G0G 2Z0
CANADA

Phone: 418-585-2612

Fax: 418-585-3953

Email: bill_jancewicz@sil.org