UNICODE - implications for the translator

Peter Kwan

Praetorius

Abstract

This paper outlines the current status of the various data encoding schemes that at present dominate the marketplace, with a brief look at how these schemes have evolved. Thereafter, the paper explores the rationale for a new encoding scheme that addresses the needs of the modern worldwide marketplace, and discusses why this is important for the translation profession.

The past...

Before the advent of the written word, or pictures, human communications was transient and probably very simple. It is not easy to prove this! With the written form, be they words... ideographs... pictures... carvings... paintings... came permanence and portability. It seems a far cry from Moses's stone tablets to today's multimedia encyclopedia, but in essence the great breakthrough was made when people moved from "speech" (transient) to "writing" (permanent).

I still recall the final scene of a film called Fahrenheit 451 (the ignition point of paper) where remaining survivors of a civilisation in the future learnt to be "talking books" in order to preserve for future generations an intellectual legacy of some kind. I preach to the converted here, as I know how fond all of you are of the written word.

Given the rich tapestry of human diversity, it is not surprising that many different forms of human communications took place. Leaving aside verbal communications, which lacked permanency and portability, we are left with two dominant forms of written communications, phonetic-based and ideograph-based, both of which evolved from picture-based written communications. There are advantages and disadvantages of each, and it is interesting to explore the graphical user interfaces (GUI) that one increasingly sees on all the latest computers, and try to understand why computer interfaces headed that way. Phonetic languages are very efficient and easy to learn and write, but lack portability and consistency (English, French, Greek, Turkish, Arabic, Hebrew ... are all different). This is because it is almost impossible to impose a consistent pronunciation or a universal usage of words. Why for example do Americans drive on parkways and park on driveways, when it would make more sense to park on parkways and drive on driveways. Ideographic languages also suffer from many problems (has anyone seen the *size* of the early versions of Chinese typewriters?) the most intractable of which is the size of the character set (over 16,000 characters in common usage). Given that there are over half a million known words in the English language, one must be thankful that there are only 26 letters in the English alphabet to learn. So a mere 26 alphabetic characters can represent so many words in the English language. It’s also astounding that just seven music notes can represent the wonderful diversity and richness of music in the world. What about the simplest representation - digital data which only has two codes? One of the great paradoxes of the modern world is that almost anything can be represented by a sequence of noughts and ones.

Data capture through history

- making primitive marks
- invention of paper and printing
- hand-written to mechanised writing
- limited (small) character sets
- digital technology
- inexpensive data storage and data retrieval

How Did We Get Here?

- Evolution of data capture - longevity, storage, retrieval considerations
- Data capture leads to knowledge dispersal and acquisition and fuels development of the human race
- The current "information age" provides the most graphic proof of this
- Information provides food for our brains
- Information frontiers are being breached

The present...

Today's Situation

- Disparate, discrete, customised data representation
- Lack of portability
- Data integrity is compromised
- Usability is poor
- Updating legacy systems is expensive
- Worldwide systems are in its infancy

Some typical present day problems

Most of the computers linked to the Internet today use EBCDIC encoding. I send an EMAIL message via the Internet to a friend in America telling him the price of the latest greatest Soccer game is £30. He says "Good, buy me a copy". Later I get a postal order for £25 from him, plus a note saying that he hoped it would cover the thirty dollars plus postage and thanks a lot for my help. I don't have the heart to tell him that I am out of pocket.

Why did this happen? It happened because the binary code for the English pound currency sign is the same as the binary code for the American dollar currency sign in the EBCDIC code page.

Similar problems arise when our European friends send us word processing files, only for us to discover on receipt that all their accented characters come out as garbage on our PCs.

The main reason why these problems have arisen is due to the limited bandwidth of the encoding systems that are common place today. These encoding systems were designed at a time when computer memory and storage was slow, and very expensive. Early telex machines for example only allowed UPPER CASE alphabetic characters and early computer printouts were also UPPER CASE only. Telex machines had a minimal set of characters to achieve simplicity of design and reliability of equipment, and early computer printers used small character sets to achieve high printing speeds (it often took all night to print out invoices that had to be sent to thousands of customers the next day). In fact, the early computer printers had enormous print chains that had scores of characters on them, but instead of the full upper and lower case characters, the print chains had many letters such as the A's and E's duplicated in order to achieve higher printing speeds.

Our own personal experiences with such things as typewriters reveal to us the limited character sets available on the typebars. Breakthroughs such as the IBM golfball typewriter gave translators the ability to switch character sets with ease by just changing the golfball typing element. True large character set capability really only became possible when all-points addressable displays and printers were made commercially available. Of course, in the publishing world, large character repertoires have always been available, but this capability was not available to you and I until fairly recent times.

With the falling cost of computer memory, and high resolution colour displays, access to a large character set is not only possible, but affordable.

A quick primer on data encoding

The mathematics of having a large encoding scheme are very simple. If you only need to store two codes, you can do that in one "bit" of computer memory. To store four codes, you need two "bits" and so on, doubling the number of unique codes possible for each increase in the number of "bits" required.

Binary data consists of bits of information that is encoded with zeros and ones (0 1)
So we can have, for example:
Decimal 0 = 00
Decimal 1 = 0 1
Decimal 2 = 10
Decimal 3 = 11 etc

One of the most well known encoding schemes is ASCII, which was originally 7-bits, with the 8th bit being used for parity checking. This allowed 2**7 (128) code points. 32 code points were used for "controls", plus another 52 for the upper and lower case "English" alphabet. Add the numbers, punctuation marks and some special characters and all 128 code points are soon used up.

In the extended 8-bit ASCII system (also known as ISO 646) there are 256 unique codes. Great! So no more problems with accented characters? Would it be so simple. No, bright programmers and designers took the opportunity to map out the more common Greek symbols, the superscript and subscript numbers, funny faces, graphic character shapes, often conveniently "forgetting" to map out the European accented characters because they were not it their everyday repertoire of characters. So there followed a proliferation of encoding schemes based on the same 8-bits. This made worldwide data exchange rather difficult.

On the other side of the world, they developed 16-bit encoding schemes in order to map out their much larger (over 10,000) character sets for Kanji, Katakana, Hiragana and Hangul. With a possible 64,000 characters, there should be no problem encoding the relatively few European accented characters ... unfortunately, this did not happen! So how do some of the large multinational Japanese companies handle French customers with accents in their names? Like most organisations, they just drop the accents. What a way to treat customers! This is changing very quickly, as multinational companies realise that to gain market share you first had to gain the hearts of your customers, and it helps to write their names correctly.

The problem was not that these various encoding schemes did not work - it was that these systems did not work well together and presented enormous problems for people when they tried to interchange data across systems and platforms.

The future…

What is UNICODE?

- a large character set - 64,000 characters
- enables the mapping of all the important languages of the world in common usage
- extendable by moving to foil 4-byte ISO 10646 standard
- an agreed industry and international standard

UNICODE is a solution that takes into consideration all the major languages of the world, making it possible to write, on a single page, a mix of English, French, Russian, Arabic, Japanese, Chinese, Korean, ...

Products like UniType allows users to write multilingual documents from within their favourite word processor, whilst more adventurous developers intent on addressing the world market can make use of the technology developed by Gamma to enhance their product offerings.

UNICODE - history

- ASCII 7-bit encoding scheme (1970's)
- LATIN 1 (ISO 646) 8-bit encoding scheme (1980's); also called extended ASCII
- ISO 10646, started in 1983, 16-bit, then 24-bit, then 32-bit
- UNICODE, started in 1987, by Joe Becker and Lee Collins of Xerox and Mark Davis of Apple
- UNICODE consortium formed in 1991, with most of key software organisations as participants
- UNICODE Standard Version 1.0 published in 1991 (Addison-Wesley)
- UNICODE Standard Version 1.1 published in 1993; this version unified with ISO 10646

UNICODE - key characteristics

- large character set (64,000)
- incorporates all the major scripts of the world
- fixed length 16-bit encoding
- unique and unambiguous encoding
- pre-composed characters
- compatible with existing encoding schemes and ISO 10646
- wide support in the IT industry
- hardware requirements no longer an issue
- key benefits to user - data integrity and increased usability

- supported by all major IT industry players - regular workshops and seminars

In summary ...

Many people still believe UNICODE to be inappropriate at the present time because

- its too expensive

- its too complex

However, the situation has changed rapidly in recent years. For example:

- 1MB of storage in 1981 cost £10
- 1MB of storage in 1995 costs £0.15
- data compression gives you more for less
- UNICODE is a large character set, but better laid out, with few compromises

Where are we today?

- Telex to EMAIL
- Books to CDs
- Static to Interactive

Recommendation

- Adopt UNICODE as soon as possible - translation industry will be one of the prime beneficiaries
- Understand how it will affect you
- Pressure suppliers to provide UNICODE compliant products and services
- Offer UNICODE solutions to vour clients

UNICODE - implications for the translator

Peter Kwan

Praetorius

Vision Statement

World peace and prosperity through better understanding

Human communications

Before the advent of the written word, or pictures, human communications was transient and probably very simple. It is not easy to prove this!

With the written word, ... pictures ... carvings ... paintings ... came permanence and portability. It seems a far cry from Moses's stone tablets to today's multimedia encyclopedia, but in essence the great breakthrough was made when people moved from speech (transitory) to "writing" (permanent).

I still recall the final scene of a film called Fahrenheit 451 (the ignition point of paper) where remaining survivors of a civilisation in the future learnt to be "talking books" in order to preserve for future generations an intellectual legacy of some kind.

I preach to the converted here, as I know you all love the written word.

Goal and Objective

- Make international communications a reality
- Make it easy for everyone

Today's Situation

- Disparate, discrete, customised data representation
- Lack of portability
- Data integrity is compromised
- Usability is poor
- Updating legacy systems is expensive

How Did We Get Here?

- Evolution of data capture - longevity, storage, retrieval considerations
- Data capture leads to knowledge dispersal and acquisition and fuels development of the human race
- The current "information age" provides the most graphic proof of this
- Information provides food for our brains
- Information frontiers are being breached

Data capture through history

- making primitive marks
- invention of paper and printing
- hand-written to mechanised writing
- limited (small) character sets
- digital technology
- inexpensive data storage and data retrieval

Some typical present day problems

I send an EMAIL message to a friend in America telling him the price of the latest Soccer game is £30. He says "Great, buy me a copy". Later I get a postal order for £25 from him, plus a note saying that he hoped it would cover the thirty dollars plus postage and thanks a lot for my help.

I don't have the heart to tell him that I am out of pocket.

Why did this happen? It happened because the binary code for the English pound currency sign on my PC is the same as the binary code for the American dollar currency sign on his PC.

Similar problems arise when our European friends send us word processing files, only for us to discover on receipt that all their accented characters come out as garbage on our PCs.

What is UNICODE?

- a large character set - 64,000 characters
  - » makes possible the mapping of all the important languages of the world in common usage
  - » extendable by moving to full 4-byte ISO 10646 standard
- an agreed industry and international standard

A quick primer on data encoding

Binary data consists of bits of information that is encoded with zeros and ones (0 1)

So we can have :

Decimal 0 = 00

Decimal 1 = 01

Decimal 2 = 10

Decimal 3 = 11 etc

One of the most well known encoding schemes is ASCII, which was originally 7-bits , with the 8th bit being used for parity checking.

This allowed 2**7 (128) code points. 32 code points were used for "controls", plus another 52 for the upper and lower case "English" alphabet. Add the numbers and all the punctuation marks and all 128 code points are soon used up.

continued ...

Well, we soon moved to full 8-bit encoding which allowed for 256 characters to be encoded. Great! So no more problems with accented characters? Would it be so simple. No, bright programmers and designers took the opportunity to map out the more common Greek symbols, the superscript and subscript numbers, funny faces, graphic character shapes, often conveniently "forgetting" to map out the European accented characters because they were not it their every day repertoire of characters. So there followed a proliferation of encoding schemes based on the same 8-bits. This made worldwide data exchange rather difficult. On the other side of the world, they developed 16-bit encoding schemes in order to map out their much larger (over 10,000) character set for Kanji, Katakana, Hiragana and Hangul.

continued ...

Great! With a possible 64,000 characters, there should be no problem encoding the relatively few European accented characters ... unfortunately, this did not happen! So how does a great Japanese company like Sony handle French customers with accents in their names? Like most organisations, they just drop the accents. What a way to treat customers!

UNICODE is a solution that takes into consideration all the major languages of the world, making it possible to write, on a single page, a mix of English, French, Russian, Arabic, Japanese, Chinese, Korean, ...

Products like UNITYPE allows users to write multilingual documents from within their favourite word processor, whilst more adventurous developers intent on addressing the world market can make use of the technology developed by Gamma to enhance their product offerings.

UNICODE - history

- ASCII 7-bit encoding scheme (1970's)
- LATIN1 (ISO 646) 8-bit encoding scheme (1980's); also called extended ASCII
- ISO 10646, started in 1983, 16-bit, then 24-bit, then 32-bit
- UNICODE, started in 1987, by Joe Becker and Lee Collins of Xerox and Mark Davis of Apple
- UNICODE consortium formed in 1991, with most of key software organisations as participants
- UNICODE Standard Version 1.0 published in 1991 (Addison-Wesley)
- UNICODE Standard Version 1.1 published in 1993; this version unified with ISO 10646

UNICODE - key characteristics

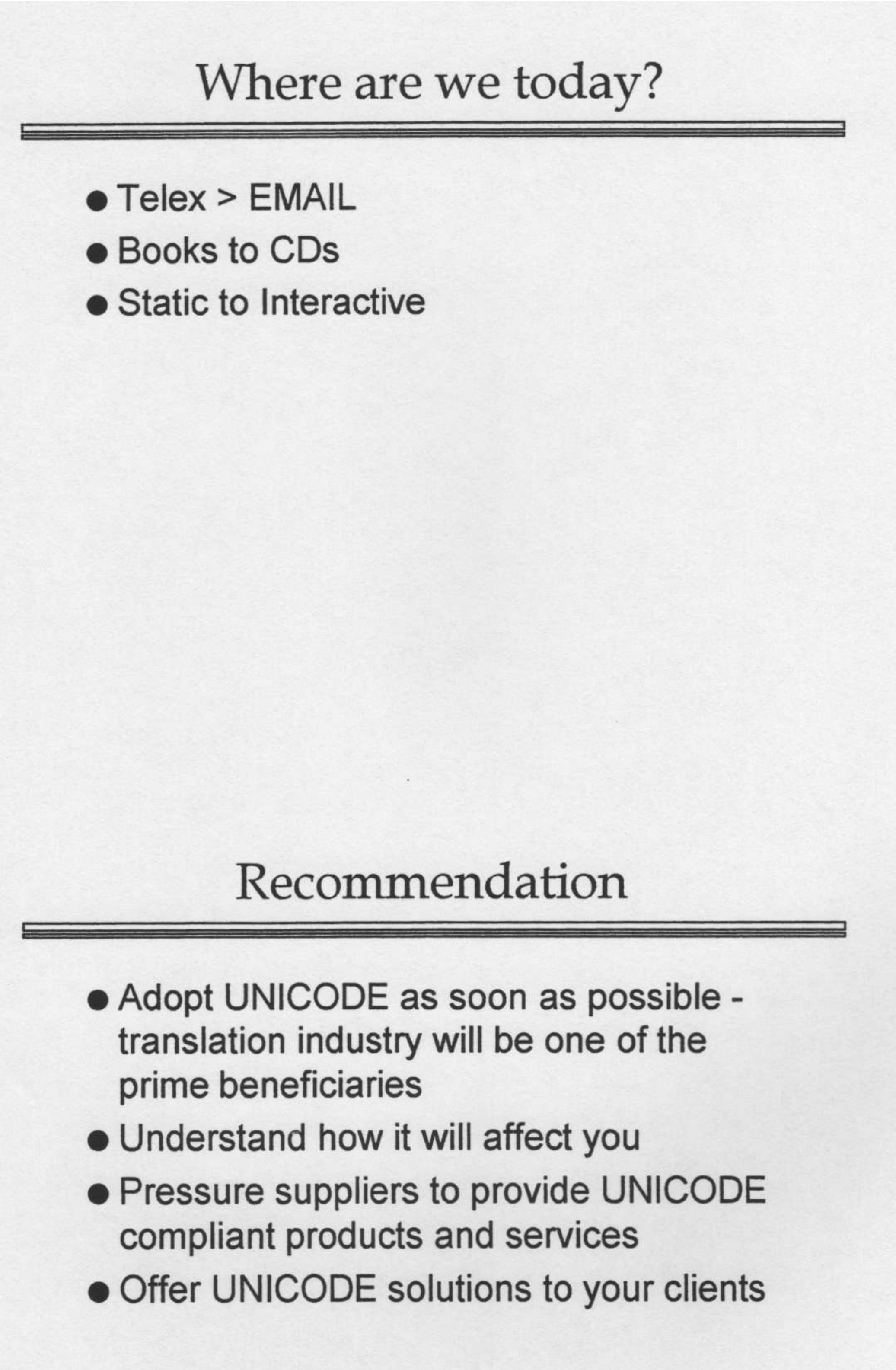
- large character set (64,000)
- incorporates all the major scripts of the world
- fixed length 16-bit encoding
- unique and unambiguous encoding
- pre-composed characters
- compatible with existing encoding schemes and ISO 10646
- wide support in the IT industry
- hardware requirements no longer an issue
- key benefits to user - data integrity and increased usability
- supported by all major IT industry players - regular workshops and seminars

Available Options

- Do nothing ... join ostrich preservation society
- Do something ... understand how large character sets such as UNICODE will affect you directly

I agree, but...

- its too expensive
  - » 1MB of storage in 1981 cost £10
  - » 1MB of storage in 1995 costs £0.15
  - » data compression gives you more for less
- its too complex
  - » no its not - its a large character set, but better laid out, with few compromises



Where are we today?

- Telex > EMAIL
- Books to CDs
- Static to Interactive

Recommendation

- Adopt UNICODE as soon as possible - translation industry will be one of the prime beneficiaries
- Understand how it will affect you
- Pressure suppliers to provide UNICODE compliant products and services
- Offer UNICODE solutions to your clients