

Unicode or Bust?: Future of Japanese computing
Nineteenth Annual Meeting of the Southeastern Association of Teachers of Japanese

Yasuhiro Omoto, University of California at Berkeley (yomoto@nihongoweb.com)

Keiko Schneider, Saboten Web Design/Albuquerque TVI (kschnei@sabotenweb.com)

0. アブストラクト

The presentation explains how Japanese script has been handled in non-Japanese localized version of operating systems and introduces the difference between double-byte and Unicode and its implications to end-users including email applications for future Japanese computing. Japanese teachers who don't work in Japan have had problems using various computer applications just because inputting and reading in Japanese has been often impossible or cumbersome. There are multiple encoding systems and that makes Japanese computing much more difficult than in Western languages. Even today many teachers use Windows 95/98/Me US versions that don't allow Japanese computing easily. The solution came with newer Operating Systems (OS). Windows 2000, XP and Mac OSX all use Unicode to deal with Japanese characters. Unicode is a result of an effort to set standards. How is it different from conventional process as double-byte? What about backwards compatibility? Can we use the worksheets we took so much time and effort to create in the older machines? And is Unicode really the solution? Mojibake problems in email are often difficult to trouble-shoot, but that is partially Unicode to blame. Your multilingual-savvy support technicians wonder why all Japanese documents and applications are not in Unicode yet. Why not? Will we eliminate all mojibake problems with Unicode?

1. オペレーティングシステムの歴史

1-1 エンコーディング：2バイト

フォントの歴史、エンコーディングの歴史はまさにオペレーティングシステムの歴史と言っても問題はないであろう。実際に、スクリーンでは、字は、左から右へという、いわゆる横書きとなっているわけであるが、これも実は、米国でコンピューターが発展してきたということを反映しているといっても過言ではない。コンピューターのスクリーンに於ける日本語表示が横書きであり縦書き表示ではないというのも実際は、米国で発達したオペレーティングシステムに日本語が載せてあるためである。中核の部分が全て横であるため、日本語表示もそれに合わせてあるわけである。

日本語表示が難しい理由の一つに文字コードの違いがあった。文字コードとは、文字などの一文字一文字に割り当てた固有の数字のことでありその数字で文字を呼び出すのだが、欧米などの言語が1バイトで表すことが出来たのに対し、日本語の場合、漢字があったためより多くの情報量が必要とされるため2バイトで表さなければならなかった。1バイトでは256文字しか表現出来なかったためである。更に、複雑なのは、この2バイトの中で色々なエンコーディングのシステムがあることである。まず、国際標準化機構（ISO）にも採用されている JIS（ISO-2022-JP）、マイクロソフトに策定され広く使用されている Shift-JIS（S-JIS）、UNIXで広く使われている日本語 EUC（EUC-JP）など様々ないわゆる標準が同時に存在する。そのため、ユーザーは、文書がどんなエンコーディングを使用しているにしてもコンピューターが即座に判定して文字を呼びだしてくれているのでエンコーディングについてはあまり意識しないですんでいる。しかし、実はその裏側で色々な、エンコーディングが、ユーザーがどれを使用しているかと意識することなく共存しているわけである。

さて、現在、コンピューターを動かす中心となるオペレーティングシステムには色々なものがあるが、日本語教育で使われている代表的なものといえば、ウィンドウズ OS とマック OS の2種類に大別されるであろう。初期の日本語教育の現場では、ワープロを使う教育者が多かったが、コンピューターで日本語を使うと言った場合には、多少の困難が伴った。

例えば、英語版の Windows 95/98/Me/NT を使用していた場合には、マイクロソフト社以外のサードパーティーが出しているソフトウェア（KanjiKit, UnionWay など）を使う以外に方法はなかった。後に、マイクロソフト社から無料で提供された純正の変換、入力システム Global IME

<http://www.microsoft.com/windows/ie/downloads/recommended/ime/default.asp>）では、日本語入力できるソフトが、マイクロソフト社製のものに限定されていたため制限があり、実際にスムーズに日本語を入出力するためには、日本語版の Windows を使用するか、後の Windows 2000/XP を待たなくてはならなかった。

一方、アップルコンピューター社の Mac OS 8.x かそれ以前のものでは、Japanese Language Kit (JLK) を別に購入する必要があったが、アップル社、真性のソフトであったため、誤動作が少なく日本語教育の初期の段階に於てマッキントッシュが広く使用される理由となった。後の Windows 2000/XP、そして Mac OS 9.x JLK は無料でシステムに付属するようになったため、日本以外の地域での日本語入力に問題はなくなった。

1-2 エンコーディング：ユニコード

初めは2バイト（UCS-2）であったが2バイトでは65536文字だけしか表すことが出来ないため、中国語、韓国語、日本語の漢字を統一して表現してしまうという問題があったため現在では4バイト（UCS-4）で定義する方式に変化した。

現在、店頭で発売されているオペレーティングシステムのWindows 2000/XP、そしてMac OSXでは、エンコーディングのシステムとしてユニコードが採用されている。もちろん、新しいオペレーティングシステムでも古いエンコーディングシステムも読めるため新しいシステム上では、古いエンコーディングシステムとの共存が可能になっている。

2. 文字化けとは何か

文字化けとは、文字が色々な理由によって意味不明な文字列に置き換えられてしまう現象のことをいう。本来の文字コードとは違う文字コードだとコンピューターが解釈してしまったり機種依存のフォントを使ってしまったためコンピューターに入っていないフォントが表示出来なかったりという理由で起ることもある。最近ではテキストをメール、インターネットを通じ転送した際に起ることが多い。この場合、トラブルシューティングは難しいが、見た目では文字化けのタイプがわかることもある。

1) 2バイト文字が的確に表示されていない例

Ç±Çíé;ÇÖi ñ {áíÇ~ÇΣÅB

この場合は、日本語のエンコーディングが何らかのが的確に表示されていないため起ってくることが多く、例えばブラウザで開けてエンコーディングを変えてやると解決出来ることが多い。

2) 2バイトのうちになくなったのがあって、組み合わせがおかしくなっている例

アれ実は日膜黷ナすけヌ見事にヤ≠/なb7したね。

この場合は、メールなどの転送の際、欠けてしまった情報があるため、起る問題で、再転送してもらうことで直ることが多い。

3) ユニコードが表示されていない

こんにちわ

??????

これは、ユニコードでエンコーディングされているのだが、正しく表示されていないということを示している。この場合、"&", "#", から始まって 5 けたの番号と ";" が、ユニコードの一文字にあたるのであるが、このような文字の並びや、クエスチョンマークの羅列を見たらユニコードであると思っても差し支えない。

2-1 文字化けの解決方法

現在ネット上で文字化けしたメールの修復をウェブ上でさせてもらえるサイトがあるので、活用してみるという方法もある。(http://www.kanzaki.com/docs/jis-recover.html)

又、他には、HTML の文書を作ってローカルでウェブサイトとしてブラウザで見るという方法もある。簡単な HTML を書いてブラウザに認識してもらおうというこの方法は、非常に効果的である。というのも、色々なエンコーディング方法が共存しているウェブで、文字コードを正しく変換するために、色々なエンコーディング方法をサポートしているウェブブラウザを使うためである。HTML を書くのはテキストが書け保存できるプログラムなら何でもよい。例えば、Windows の場合、NotePad でいいわけであり、マッキントッシュの場合は、SimpleText (OS9), BBEdit (OSX) <http://www.barebones.com/products/bbedit/>などでいいわけである。

書き方であるが、文書を.html の拡張子を入れたファイル名で保存し、初めに<html>終わりに</html>というタグを入れればそれで出来上がりである。それをブラウザを使ってローカルにファイルを開け、正しいエンコーディングを選べば良い。次に HTML の例を載せて置くので参考にして頂きたい。

例

<html>

Ç±Çíé;ÇÕi ñ {áÍÇ≈ÇΣÇØÇ«â©éñÇ...É_ÉÁÇ...Ç»ÇÈÇ<ÇμÇΩÇÀÁB
±Çíé;ÇÕi ñáÍÇ≈ÇΣÇØÇ«â©éñÇ...ÉÉÁÇ...Ç»ÇÇ<ÇμÇΩÇÀÁB

</html>

以上のように、文字化けしてしまったテキストを<html></html>で挟めばよいわけである。

2-2 文字化け：サーバーのせいで直せないもの

以下のような例は残念ながらどのような手を尽くしても変換することが出来ない。

1)

_O Xæ ¶

± ñ É ¿ [B ¼ Ä * á w] | à Ä · BJHL @SIG | - [e B O Ä · * A ° Ð Q Á µ ½ c Æ v Á Ä c Ü · } Ä A Ç □ ¼ æ ë µ - ` è c µ Ü · B

| à

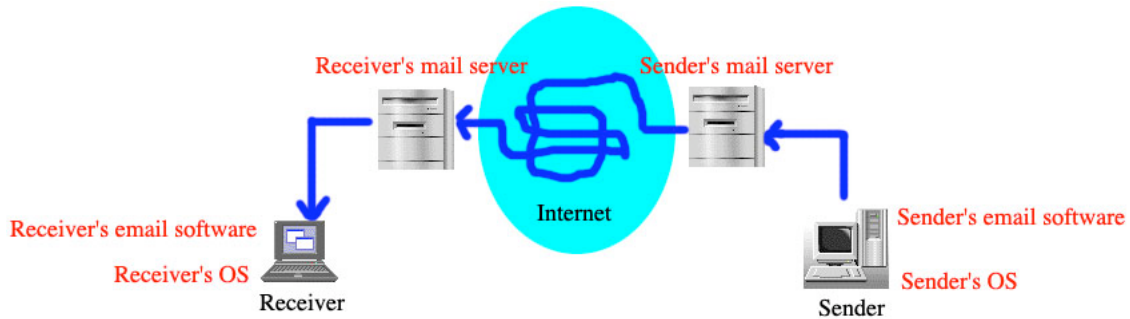
2)

\$B2<5-\$N\$H\$*\$j!"Bh (B6 \$B2sG/<!9q:]Bg2q\$R3+-\$-^\$9!#8&5fH/l= (B(\$B8}F,!"]%9%?!< (B) \$B\$*\$h\$S (B \$B%7%\$s%] %8%&% ` \$N4k2h\$Rjg=8\$7\$F\$*\$j\$^\$9\$N\$G!"@'Hs\$41~Jg2<\$5\$K\$h\$&!"\$*4j\$\$\$\$\$? \$7 (B

このような場合は、サーバーで文字化けが起ってしまっているので、文字化けを直すことが出来ないのである。このような文字列を見た場合には、違う手段でもう一度送り直して貰わなければならない。

3. 電子メールの文字化け：誰の責任なのか。

簡単に以下の図を見て欲しい。電子メールの場合、どうしても文字化けが起った場所を特定するのが難しい。それは、以下のように文字化けを起すポイントが幾つか考えられるからである。送り手の OS、送り手の電子メールソフト、送り手のサーバー。そして中間地点のインターネット。受け手のメールサーバー、受け手の電子メールソフト、そして受け手の OS と、普段我々が簡単にメールの送受信として済ませてしまっている裏に、これだけの文字化けの要素が含まれているわけである。



このように文字化けの原因が色々考えられる場合、ありとあらゆる可能性が存在する。例えば、ユニコードで送られた文書が読めないとすると、受け手の OS がユニコードが扱えない場合文字化けが起ってしまうということがある。又、その場合、OS がユニコードを扱えても電子メールソフトがユニコードが扱えなければ文字化けが起ってしまう。更には、メールサーバーに問題があるかもしれないし、その中間地点のインターネットの中継で情報が欠落してしまう可能性もある。又、OS が違えば機種依存文字によってフォント表示が出来なくなる可能性もある。それだけではなく、全く存在しないフォントを特定して送ってしまった場合など、置き換えが自動的に起らずに文字化けしてしまうということさえあるだろう。例えば、日本語のフォントがないコンピューターに日本語のテキストを送っても表示が出来ないというのがその簡単な例である。斯様に、文字化けは複雑であるため原因を特定するのは非常に難しい。

4. 日本語コンピューティングの将来：ユニコードになるのか

知らないうちに新しい OS でユニコードを使っているユーザーが増えている。ユニコードだと他言語が使用出来るという利点があるとはいうものの、ソフトウェアの中にはユニコードに対応していないものも多い。

又、ユニコード自体が欧米の業界で策定が進んでいるということもあり、反対している人々も多い。最初の頃、中国語、韓国語、日本語の漢字を同じように処理してしまうというのに反対していたのはその良い例であろう。

しかしながら、ユニコードにも利点が多い。例えば、基本的にジャバで書かれたアプリケーションはユニコードをサポートしているケースが多い。そのため、ジャバのプログラムで日本語が通るという可能性が高いためジャバのプログラマーが、特別に日本語などにプログラムを対応させなくても日本語が使えるという有利な点があるのも事実である。とはいえ、ユニコードになったからといって日本語フォントを入れなくてもいいということではなく、やはり日本語のフォント

がないと日本語表示は出来ないのである。そして、北米で出荷されたコンピューターがそのままの形で日本語の読み書きが出来るようになっていることはまずないため、日本語教師は日本語コンピューティングに関して、テクニカルサポートの担当者と一緒に勉強していかなければならないであろう。

5. ユニコードと日本語教育の現状

ユニコードへの移行は新しいOSの裏に隠れて気がつかれないまま進行している。そのため、実際に自分がユニコードを使っているということに気がつかずにいるユーザーが大半である。新しいOS、新しい電子メールソフトで送られたメールが古いOS、古い電子メールソフトで受けた場合読めないという可能性が起るのには避けられない。全てが一瞬でユニコードに移るというわけではなく、この変化は少しずつ起っていくと考えられるので、後何年かはこのような混乱が続くことが予想される。現在、アメリカで30%程度の人々がまだ、Windows 98を使っていると言われている。一般社会でも移行に時間がかかっているため教育の現場では、特に金銭的余裕のない教育機関に於ては古いOSを使い続けなければならないという問題が生じてくるであろう。そのため今後何年もの間、文字化けが起るという可能性は否定できない。

しかし、古いOSだからといって全然、ユニコードが読めないというわけではなく、問題が分かっていたら対処の方法は、前に書いた通り色々あるので、まずは問題の認識が重要であると考えられる。

6. 学習者と日本語コンピューティング

以前、シュナイダー、深井、尾本が、アメリカでの教育者と教育機関のコンピューター事情を調べた結果と尾本が調べた学習者のコンピューター事情を比較した結果、教育者よりも学習者の方がITスキルが高いケースが多いということが分かったが、日本語コンピューターに関してはこの限りではないという結果が出た。つまり、どんなにコンピューターに習熟している学習者でも日本語に関しては指導してやらなくてはいけないのである。そのため、教育者も、もし、コンピューターやインターネットを活用した日本語教育を行うという場合には、日本語コンピューティングについてある程度知っていなければならない。又、日本語コンピューティングをサポートするためのテクニカルサポートの担当者もある程度

の日本語コンピューター環境についてしらなければならないであろう。そのため、教育者やテクニカルサポートの担当者に対しての IT リテラシーワークショップなども必要ではないかと考えられる。

7. ウェブ上の日本語コンピューティングに関する情報サイト

ワークショップ等に参加して学習者に日本語コンピューティングを教えられるスキルを身に付けるということをしなくても、基本的なことであれば、学習者にウェブサイトを指し示すだけでいいということがあるだろう。例えば、日本語のフォントなどのインストールは、英語、又は簡単な日本語で書いてあり、やりかたが、書いてさえあれば学習者自身が、自分のコンピューターにフォントをインストールして日本語環境を整えるということは可能であろう。

日本語環境を整えるためのサイトとして、例えばウィンドウズでは日本語オーケー！ドットコム (<http://www.nihongo-ok.com/>) があり非常に便利である。マッキントッシュでは、例えば NihongoWeb など (<http://www.nihongoweb.com/>) が便利である。又、NihongoWeb や Keiko Schenider's Bookmarks (<http://www.sabotenweb.com/bookmarks/>) などには、日本語コンピューターについての他の有益な情報や、サイトへのリンクがあるので、それを活用するといいいのではないか。

8. 日本語教師のための IT リテラシーと情報共有

以上のような日本語コンピューターを巡る問題の根は深い。日本語コンピューティングを巡る様々な問題を解決するためには、やはり問題を共有し更にその解決方法を共有するということが必要だと考えられる。そのために先生オンライン (<http://www.sabotenweb.com/bookmarks/about/senseiOnline.html>) などのオンラインコミュニティで教育者が問題、解決法、アイデアなどを共有して行くことが必要であろう。

参考資料：

インターネットメールの注意点

<http://www02.so-net.ne.jp/~hat/imap/cover.html>

漢字コードと文字化け

<http://www.kaiteki-net.com/nettrouble/mojibake/>

CJKV Information Processing by Ken Lunde, O'Reilly

<http://www.praxagora.com/lunde/cjkv-ip.html>

参考文献：

Omoto, Y. (2002). "Utilizing the Internet for classroom teaching". In Tohsaku ed. Proceedings of The third International conference on computer assisted systems for teaching and learning Japanese CASTEL/J 264-268

Schneider, K. (2002) "Implications on Teaching Japanese with the Internet". In Tohsaku ed. Proceedings of The third International conference on computer assisted systems for teaching and learning Japanese CASTEL/J 260-263

Omoto, Y., Schneider, K., Fukai, M., (2003) "Do Japanese teachers take advantage of technology?: A survey on the use of technology in Japanese classes in Northern California" The Breeze No. 29, 10.