Unicodeは好きですか？

和田英一／（株）富士通研究所

日本の反対を押し切って規格化されたUnicode（ISO 10646のBMP）は多くの欠点がある．それにもかかわらずJavaなどと一緒に我が国に上陸しつつある．この際もう一度問題点を再認識し状況を把握しよう．そして悪いde facto standardがはやるのを防止しよう．

はじめに

SIGPLAN Noticesを読んでいたら，Alan Kayが書いていた[1]．

「われわれはGreshamの法則を避けねばならぬ．Greshamの法則とは経済学で悪貨は良貨を駆逐する，というものである．悪いソフトウェアの問題点の1つは，良いソフトウェアより悪いソフトウェアの方が作るのが容易であることだ．人々はunderstanding-orientedであるよりはgoal-orientedなので，この悪いソフトウェアを使い，除去するのに何十年もかかるde facto standardを作ってしまう．悪いソフトウェアを挙げようとは思わないが，1つだけいうと，1965年に悪いアイディアだといわれていたMS DOSは，それにもかかわらず，very unsophisticatedな人達が作り出し，even less sophisticatedな人達に使わせた．これがGreshamの法則である．**de facto standardは将来に前進しようとするわれわれを殺すものだから，この分野の専門家として，これらと戦わなければならない．**将来のソフトウェアに起きることの多くは，その意味するところのunderstandingの結果として起きるのだと確信する．」（太文字は和田）

これから議論しようとするUnicodeはまぎれもなくgoal-orientedで，悪いde facto standardを広めようとしているのである．

そもそもUnicodeはISO10646という規格（JIS X 0221に対応）のBMP（後述）におかれた文字集合である．ISO10646は筆者がISO/IEC JTC1/SC2の国内主査をしていた1984年に，情報処理用の文字集合として規格化が始まったものである．最初は2オクテットの固定バイト長の規格として，コード表の構造から検討が始まり，1987年には第1，第2オクテット共0x20-0x7f，0xa0-0xff（0xは16進を表す）の範囲にのみ図形文字を配置する；1990年には表意文字の領域は各国の国内規格をそのまま入れる；との合意ができていた．

それが米国のUnicode派が，合意をひっくり返し，日本，中国，韓国の漢字を統合したHCC（後述）を入れる提案をして規格を通してしまったのである．日本はとてもこれでは日本の内外を問わず使えないと反対した．

反対の理由をいくら説明しても，自社企業の世界戦略だけに目がくらんでいる米国人のインテリジェンスでは理解の範囲を超えているらしく，まったく分かってもらえなかった．最近，Javaと関連したUnicodeの記事を読んだが，ますます大変なことになっていて[2]，いわんこっちゃない，といいたい状況である．太田君の本にも書いてあるが[3]，本稿でも，Unicodeがなぜ使いものにならないかを，もう一度整理して述べたいと思う．

制定の経緯と問題点

前述のようにISO/IEC JTC1/SC2では，1984年頃から世界中の文字を配置したコード表が欲しいということで，WG2を新設して作業を開始した．WG 2では最初このコード表の基本構造を議論した．8ビットのオクテットを2オクテット使った65536個のコード点をすべて使うB案（英国の提案で後述）に対し，各オクテットとも制御文字の場所を避け(95+96)×(95+96)=36481個のコード点だけを使うA案（米国の提案）とがあり，長い間その得失の議論が行われた．どちらの案でも日中韓の漢字（およびハングル）部分は各国の国内規格の表を（必要があれば平行移動して）収めるのもであった．B案の方が多くの文字が入るがやはりA案のようにC0，C1の制御文字の場所はあけておきたい；B案はA案より漢字コード表の移動が厄介；収めるべき文字が予想より増えてきた結果2オクテットでは無理になった時，オクテットの延長にはA案が楽；ということで1987年3月，SC 2のベルリン会議で投票の結果，9対1（棄権3）でA案で作業を進めることになった．その後図形文字の充填作業が進行し，DP 10646は第2版まで進んだ．

DP 10646は4オクテットが標準だが，デフォルト

で第1，第2オクテットを32に固定し，第3，第4の2オクテットだけで使うことができる．これをBasic Multilingual Plane（BMP）といい，ここには上述の36481個の図形文字が入る．これを4つの4分区（00，01，10，11）に分け，各4分区で上の16行に95～96文字に収まるコード表を入れ（A00，A01，A10，A11），残り79～80行に日中韓の国内規格の漢字（およびハングル）部分を入れる（I00，I01，I10，I11）のである．日本のJIS X 0208-1983はI00，中国のGB 2123はI10，韓国のKS C 5601はI11となっており，I01は特別な使い方をすることになっていた．

ところが中国には簡体字と繁体字と合わせて4分区大のコード表が6つあり，これをBMPに入れたいという無理な希望を持っていた．そのため，JIS X 0208の解説にある「字形の違いがわずかであると認めるものはただ1つの符号に合併する（解説表6参照）」を盾にとり，日中韓の同じ形の文字はBMPに1つ入れるだけにしようという提案をしてきた．これをHCC（Han Characters Collection）という．

これにヒントを得たのか，米国ではXeroxのJoe Becker，AppleのLin Collinsを中心として，2オクテットですべてを入れるUnicodeの計画が始まり，これがHCCのアイディアを採用して日中韓台の規格をマージして入れたUnicode0.0の案ができ，米国よりDP10646に代わるものとしてUnicodeを採用せよとの提案となって，DP 10646のDIS化に障害が起きた．Unicodeは:

• C0，C1の制御用の部分にも図形文字を入れた;
• アクセント記号のある文字は合成文字とした;

などの一般的問題はあるものの，さらに我が国にとっては，漢字部分を日中韓台の漢字をマージしたHCCとしたことで，重大な関心を持たざるを得なかった．

WG2はその後の会議で日，中，韓，米間でHCC是非論が続いたが，結論をみるに至らなかった．大体の傾向は，日本はHCC，Unicodeに反対，米，中は賛成，韓国は考慮中ということである．ここでは結論が出ないので日中韓のエキスパートが1990年，ソウルでこの件に関する会議を開き，DP 10646はその結論に従うということになった．

そのソウル会議に向けて用意した日本がUnicodeないしHCCに反対する理由はおよそ次のとおり:

• 漢字には義，音，形があり，義が一番重要と思われるのにHCCは形にしか注意しない．
• Unicodeがマージした各国の国内規格の文字集合はそれぞれの国が独自の仕方で選んだものであり，その単なる混合は国際性がない．
• それぞれの国内規格は文字の追加が必要になる可能性があるが，Unicodeのメインテナンスは困難である．
• 各国の文字は一見同じようでも，日本と中国で微妙に違っていたりし，同一コードに対し，活字をかえなければならない場合があり，Unicodeにタグをつけるならば DP 10646の方が便利．
• Unicodeの漢字配列は他の国の同形文字の存否に依存し，文字の探索が困難である．Unicodeのコードは JISのそれとまったく違うので，コード変換には大きい変換表が必要である．日本はすでに大量の日本語のファイルを持ち，いまさらコードを変えるわけにはいかない．

2月末のソウル会議では結局以前の通りの平行線を辿り，DP 10646はとりあえず変えないことになったが，韓国はHCCを推進する，あるいは推進できるかどうかを検討するためCJK-JRG（中日韓Joint Research Group）を設置することを提案し．各国は3カ月以内に参加する，しない; 参加ならメンバを韓国に通告する; ことになった．

ソウル会議では日，中，韓，台が今後コード表に入れたい文字数の調査があり，

• 韓国はKS C 5601にあと8000字のハングル，22000字の漢字を;
• 日本は6000字の補助漢字集合を;
• 中国は各7000字の簡体字コード表を2つ，それらとGBに対応する繁体字のコード表を3つ; 台湾は70000字を;

それぞれ追加したいというので，これだけでUnicodeは破綻している．それでも韓国や中国は検討会を持ちたかったのか．

SC2では，ワシントンDCにおいて，1990年4月4日よりWG2会議，9日より総会を予定していた．Unicode問題でまた多少はもめるのではないかと予想されたが，2月のECMA TC1会議はUnicodeに反対，DP 10646に賛成の態度を決めた．3月のANSIの会議でもJISのコード表はそのまま入れようという方針になり，ISOではUnicodeはいったんは流れる見通しになった．しかし，米国ではUnicodeに基づいた製品も出回るかもしれず，楽観できない状況にもあった．一方ISOがDP 10646を採用すればCJK-JRGは宙に浮く可能性もあった．

DIS 10646はUnicodeとの併存を避けたいという思惑とUnicode派の組織的圧力により反対多数で否決された．その後UnicodeとマージしたDIS 10646-1.2がそれまでの検討内容と似ても似つかない形で登場し，これがISとなった．

その後の展開は小林さんの解説を見て欲しい[4]．

上に列挙した問題点以外にも，Unicodeは国内規格にない文字もCJK Compatibility Ideographs（CJK互換表意文字）などに採り入れたりして，ひたすら汚い方向へ進み，とても国際規格という顔ができる状態ではない．

なお漢字の統合に関しては，文献5）の409ページ

に，『中国の簡体字との関連』として，

> 文章の表記に漢字を使用している国々には近い所で，中国，韓国，台湾などがある．新しい漢字表を考えるにあたって，これらの国々の漢字対策についても知った上で，国語審議会の検討をしていくことにした．中国，韓国における漢字使用の現状，教育における文字指導の問題，漢字施策等について調査団が派遣され，その報告が審議会でなされた．
> 特に，字体の簡略化については，字体を検討していることでもあり，新しく採用する字体として，中国で使用されている「簡体字」と共通にする方向を考えるか否かが話し合われた．しかし，審議会のおおかたの意見は，字体を簡略にしてゆく基盤や国情が全く違うので，中国の簡体字の字体については顧慮する必要がないということであった．
> なお，審議の過程の中で，日本と中国の略字の共通化について国会で質問（昭和52・10・18，参院予算委）があり，文部大臣が次のように答弁している．
> 「現在国語審議会で漢字表及び字体表の審議が行われており，字体の審議に関連して審議会委員を中心とした調査団を中国に派遣して中国の文字改革の実情を調査したこともある．この問題は両国の文字改革の方向，文字使用の実態に相違があり，両国のそれぞれの文化に根ざした経緯もあるので，今後とも文化庁国語課で中国の文字改革の情報等を収集していくとともに，当面は国語審議会の議論を通じて，検討，判断を続けていきたい．」
> 国語審議会としては漢字表委員会で検討したが，結果としては前記のようなことになった．
> したがって，常用漢字表の字体は，中国の簡体字とは関係なく考えられているものである．

という記述がある．

ではどうするか

どう考えてみても統合した漢字コードは無駄である．さらに言語情報を使うというに至っては，何をかいわんやである．それに我々にはまったく用のないハングルが1万1千字以上もあるのもコードポジションの浪費でしかない．またBMPはやがて満杯になり，別の面を使うことになるらしい．

一方，テキスト処理のプログラムをある程度国語に独立にしようとすると，制御文字は共通にしておきたい．一応ISO 2022の枠組みの中でも，G0にASCIIをおけば，制御文字は固定になる．ただASCIIは1オクテット，漢字は2オクテットというところは都合が悪い．

したがってすべてを2オクテットにし，1オクテットのASCIIや2オクテットの漢字はその枠組みに入れて処理すれば，システムを書く方もシステムを利用する方も楽になるに違いない．こうなるとベルリン会議の前に議論していたB案というのもなかなかよかったという気がする．情報交換にはISO 2022のエスケープを使って文字集合を識別し，計算機内ではB案の枠組みで文字情報処理をするのである．

B案は記憶が多少あやふやだが，

領域	コード数	使用目的
0x0-0x3f	64 (32×2)	C0 + C1
0x40-0xff	192 (96×2)	ASCII + 8859-1
0x100-0xfff	3840 (96×40)	40 single byte sets
0x1000-0xffff	61440 (96×80×8)	8 double byte sets

（ダブルバイトは16区-95区が本質的とすると80×96コードポイントで十分という考え）という構造になっていた．

おわりに

悪いde facto standardを作ろうとしているUnicodeはなんとか退治しなければならない．

いったん入り込むと撲滅は困難至極である．米国にたくさんの使用例，存在例があるからといっても歓迎は禁物で，拳銃，麻薬，エイズなどと同様，Unicodeも水際で撃退する断固たる決意が必要である．

参考文献
1) Bergmann, S. : History of Programming Languages Conference II, SIGPLAN Notices, Vol 31, No.11, pp.9-20.
2) 凪間一洋: Java最前線7 Javaの国際化と日本語処理 bit, Vol.29, No.12, pp.91-101.
3) 太田昌孝: いま日本語が危ない－文字コードの誤った国際化, 丸山学芸図書 (1997).
4) 小林龍生: マルチリンガル文書と文字情報処理2 文字コードにとっての真の国際化とは, bit, Vol.29, No.11, pp 48- 54.
5) 藤原　宏: 注解常用漢字表 新しい国語表記, ぎょうせい (1981).

（平成9年12月11日受付）

インタラクティブ・エッセイとは

インタラクティブ・エッセイは，著者の主張をベースとして，会員読者がインターネットを使って議論に参加できる新しい試みのコーナです．著者主張の脱稿時点から学会ホームページに掲載を開始し，コメンテータの異議，それに対する著者反論と議論が進む過程を逐次ホームページに掲載します．電子メールで意見をお寄せくだされば，編集して著者とコメンテータに取り次ぎます．ページ数と印刷時期の制約により学会誌に掲載できなかったものは，編集して学会ホームページに掲載します．以下のURLをご覧ください．

http://www.ipsj.or.jp/magazine/interessay.html

では，文字コードは好きですか？

戸村　哲／電子技術総合研究所

文字コードや漢字の情報処理が今のままで良いのかという問題は，日本文藝家協会[1]が問題提起を行って国語審議会に要望書を提出したり，「電脳文化と漢字のゆくえ」[2]という本が出版されるなど，身近な話題になりつつあります．また日本工業標準調査会の国際部会は答申「今後の我が国の国際標準化政策の在り方」の中で，重要分野の1つとして「多言語情報処理」をあげており，文字コードを含む多言語情報処理技術の重要性が認識されはじめています．

さてUnicodeが好きかと聞かれれば，好きでも嫌いでもありません．でもUnicodeコンソーシアムと標準化団体（ISO，日本工業標準調査会，日本規格協会）を比べてみましょう．たとえば，Unicodeの規格書[3]に比べるとISO 10646-1やJIS X 0221の規格書は2，3倍の値段がつけられています．しかも文献3）にはさまざまな解説が添えられており，役に立ちます．またUnicodeのホームページ[4]では他の文字コード規格との対応表など有用な情報が得られます．Unicodeは今の文字コードを集めたものであり，多くの文字コード規格を調べることができます．つまりUnicodeは便利です．

次に統合漢字について考えてみます．Unicode（またはISO 10646-1のBMP）の「統合漢字」もJIS X 0208の「字体包摂」も漢字を文字コード化する枠組みは基本的に同じです．しかも原規格分離漢字規則により，国内漢字規格（JIS X 0208およびJIS X 0212）にある漢字はすべてUnicodeの統合漢字にあります．コード変換は必要ですが，統合漢字に限れば，国内漢字規格に関してUnicodeは新たな問題を導入していません．

Unicodeは従来の文字コード技術を再検討するための一里塚です．漢字の問題にしても，たとえばUnicodeの統合漢字をたたき台として，網羅的に問題点を列挙し，具体的対策を検討することが重要です．具体的な議論を積み重ねることで，新しい多言語処理技術が生まれてくると期待しています．ぜひ楽しい議論をやりましょう．

参考文献
1）http://www3.mediagalaxy.co.jp/bungeika/
2）平凡社編：電脳文化と漢字のゆくえ－岐路に立つ日本語，平凡社，(1998).
3）The Unicode Consortium: The Unicode Standard, Version 2.0, Addison-Wesley (1996).
4）http://www.unicode.org/

（平成10年2月10日受付）

これが反論か

和田英一／（株）富士通研究所

戸村からやっと反論と称するものがきた．師弟関係で遠慮する戸村とは思えないから，ずいぶん時間がかかったのはやはり反論の種がなくて，苦労していたものと思われる．

考えてみれば日本中，Unicodeが好きな人はいないわけだし，好きというのは余程どうかしている人だから，この反論を依頼されたのは貧乏くじを引いたということで，同情には値するが．

規格書の値段は本質的でない．「マルチリンガル環境の実現（INOCODE編）」を安値で出版すればよいのだ．「統合漢字」と「字体包摂」は同じ枠組みで，JISの文字がすべてあるから問題ないとは甘い考えで，そのため言語情報なしでは処理できなくなっている．また以前戸村には伝えたごとく，国内規格で包摂したものを再度Unicodeで包摂するのは，包摂の趣旨から認めるわけにはいかない．

Unicodeは出発点から間違っているのだ．苦しみの結果の反論かと思うが:

- 規格としてポリシーが認められない．
- 個々の利用者からみて不用な文字が多すぎる．
- 国内規格とのコード変換が必要．
- 情報交換にもコード変換が必要．
- 国内規格の改訂に追従が困難．
- Unicodeが1面に収まらなくなり，実現が困難．
- 言語情報の追加で，処理が複雑化．

など重要な問題点を戸村がすべて容認しているとすれば，電総研の研究官からしてそのような認識では，困るとしかいいようがない．

「Unicodeの統合漢字をたたき台とし」たのでは，なにも生まれない．

（平成10年2月12日受付）

Unicode批判批判

萩谷昌己／東京大学

●Unicodeでもいいじゃん

和田先生の文章に，「考えてみれば日本中，Unicodeが好きな人はいないわけだし，好きというのは余程どうかしている人だ」とありますが，ここは1つ，悪役を買って出て，Unicode賛成の意見を述べさせていただきます．

そもそも日本語なんていうものは，借り物だらけの言語ですよね．特に，書き言葉は．この文章の中でも，Unicodeというように，ローマ字も使っているし，当然ながら，漢字という中国語の文字が混じっているわけです．日本人は偉そうに漢字を使っていますけど，すべて中国からの借り物ですね．そう考えると，「てにをは」を書くためのひらがなとカタカナさえ入っていれば，それは日本語と考えてもいいんじゃないですか．ですから，Unicodeは日本語コードだと思うこともできるんです．Unicodeを日本語コードとして採用すれば，日本語は，漢字やローマ字だけでなく，世界中の文字を借りることができるわけです．

多少，漢字の形が違っていてもいいじゃないですか．その代わり，ハングルが使えます．ソウルと書かずに，서울と書くことができるわけですね．日本語の中にハングル文字がいつも自然に入っている．たとえば，韓国の大統領の名前とかは，漢字ではなくハングルで書いてある（ひらがなでふりがなが振ってあってもいいですけど）．Unicodeを採用したおかげで，日本人が皆，ハングルを読めるようになったりしたら，面白いじゃないですか．

統合漢字の中に常用漢字が実質的に含まれているのなら，それで現代日本語は十分に表現できるわけですし，そうならば，メリットの方がはるかに大きいのではないかと思うのです．Unicodeが世界的に採用されれば，日本語の交じった文章を，何の苦労もなく世界中の人に読んでもらえるわけです．また，アジアの情報も非常にスムーズに入ってくることができるようになります．最初は影響は少ないかもしれませんが，generationを経るにつれて，その影響ははかりしれないものになると思います．

Unicodeを採用することによって，中国の呪縛からも西欧の呪縛からも，自由になれるような気がしませんか．

●攘夷 vs. 開国

前節で，私の個人的な思いを述べさせていただきました．かなり斜に構えた意見の展開だったかもしれませんが，基本的には，「国際性」という論点に落ち着くと思います．

芝野耕司氏は，文献1）において，攘夷と開国という言葉を使っています．また，Unicodeができた理由は，（国際的な文字コードの体系化に消極的だった）日本のサボタージュのせいであると言っています．きっと彼に言わせれば，和田先生も攘夷派でしょう．「日本語という純粋な言葉を，素性のわからない怪しい外国語とごちゃまぜにされてたまるか」という思いは，日本人ならば誰にでもあるはずです．

和田先生があげているUnicodeへの批判は，どれもあまり説得力がありません．

- 統合文字は形しか注意していない: では，意味に基づいて統合ができるのですか．芝野氏がいうように，もともと漢字には多義性があり，意味によって統合などできるはずがありません．
- 単なる混合は国際性がない: そもそも，日中韓にまたがるような国際的な言語活動など，現在のところ存在しないわけで，統合文字を作ろうとしたら，現在の各国における言語活動に依存せざるを得ないじゃないですか．国際的な言語活動を可能にするためにも，まず，文字コードの統合が必要なわけです．
- 国によって文字が微妙に違う: 日本語においても文字の揺れはあります．これは我慢すべきことです．外国のレストランで日本語のメニューが出されたとき，へんちくりんな漢字が書いてあったりしますが，あういうものを許容する感覚が必要ではないでしょうか．
- メンテナンス・検索・コード変換の困難さ: これらの点はいちゃもんとしか思えません．計算機パワーの向上を考えれば，変換テーブルなど些細なことではないですか．

以上のような説得力のあまりない批判は，どうしても，その背景として攘夷的な感情を考慮に入れないと，理解できそうにないのです．どうでしょうか．

参考文献
1) http://spin.asahi.com/paper/aic/clipping/9802ks/index.html

（平成10年3月18日受付）

～ 議論の続きは，次のURLをご覧ください．http://www.ipsj.or.jp/magazine/interessay.html ～