平成23年度　国立情報学研究所　市民講座　第4回
「インターネット時代の文字コード―漢字コードの迷信を打破する！―」

講師：宮澤　彰
（国立情報学研究所　情報社会相関研究系　教授）

◆　講　義　◆

こんばんは、宮澤です。
総研大という大学院で講義するときはせいぜい数人で、一人ということもよくあるぐらいです。
それに比べ、これだけ顔が並んでいると、普段とは違って少し緊張します。

さて、「インターネット時代の文字コード―漢字コードの迷信を打破する！―」というタイトルを付けております。

・スライド2「質問1」

最初に質問。別にテストではないので気楽に答えてください。

「日本語は、世界で一番難しい言語だ」。
Yesと思う人？（挙手）
Noと思う人？（挙手）
はい、どうも。
半々ぐらいですか。

・スライド3「質問2」

「今の文字コードには漢字が足りない」。
Yesと思う人？（挙手）
Noと思う人？（挙手）
はい、どうも。
なかなか圧倒的とはいかないものですね。

・スライド4「質問3」

「漢字はおよそ何文字あるか」。
これはなかなか難しいですが。
2万字ぐらいだと思う人？（挙手）
5万字ぐらいだと思う人？（挙手）
8万字ぐらいだと思う人？（挙手）
はい、ありがとうございます。
先に予習しちゃったりすると・・・。

正解は後で言います。

・スライド５「質問４」

「漢字には意味がある」。
Yes と思う人？（挙手）
No と思う人？（挙手）
はい、どうも。
正解は、最後までちゃんと聞いていてくだされば分かります。

・スライド６「文字化けしたウェブページ」

最初に、こういう話でおなじみだと思いますが、文字化けはコンピュータを使っていると誰でも経験があると思います。ｓ例えばホームページが文字化けして、読めなくなってしまう。
この対処の仕方なども面白い話ではありますが、それは今日の主題ではありません。

・スライド７「コンピュータ内部での文字表現」

なぜこういうことが起こるかについて、少し技術的なことになりますが、最初に簡単にお話ししておきます。
コンピュータは「1／0」、オン／オフの二進法でしか物事を扱っていないということはよくご存じだと思います。それを普通は１と０の列で表しますが、最近では、１と０の一けたをビットといいますが、これを８ビットずつ区切ってファイルに入れるようになっています。
そのファイルの上で、例えば「11100110　10111001」とあると、これをコンピュータがどう解釈するかはソフトウエア次第なのです。
文字列だと思ったときに、エンコーディングの方式を教えてもらわないと、どういう文字だかよく分からない。

ここに灰色の背景で書いてあるのが、エンコーディングの名前です。
Shift-JIS は、皆さんよく聞くと思います。
EUC-JP、あるいは ISO-8859-1 など、いろいろな種類があります。

「11100110　10111001」のビット列を Shift-JIS だと思うと、「貉（むじな）」になります。
その上に書いてある「35977」が内部コード、つまりこの文字に付けた中での番号です。
一方、全く同じビット列を EUC-JP だと思ってみると、「羚（かもしか）」という字になります。
それを ISO-8859-1 という主に西ヨーロッパで使われているエンコーディング方式で見ると、１文字ではなく、２文字になります。
最初の「11100110」の文字が、「a」と「e」のくっついたような、「æ（エーイーダイグラフ）」という文字で、ラテン語などで使います。
その後ろの「10111001」が、「1 の上付き数字」といいます。
このようなものとして解釈されてしまいます。

このエンコーディングを間違えると、先ほどのような文字化けが起こるのです。

・スライド8「エンコーディング」

エンコーディングがどうしてそんなにたくさんあるのかは、歴史的事情によるとしか言えないのですが、どれぐらいあるかというと、百何十種類です。
ブラウザなどでエンコーディングの種類が選べるようになっていますが、そこには140幾つ、ずらりと全部出てくることはありませんが、それでも数十種類は普通選べるようになっています。
それに別名があります。
ISO-8859-1は、別名ではウエスタンヨーロッパ、西ヨーロッパという名前なので、それも勘定すると400、日本語を入れると500、600、700という数になってしまいます。

割合とよく聞くエンコーディングの例として、ASCII、BIG5があります。
BIG5は台湾、GBKは中国、EUC-KRは韓国です。
LATIN1は8859-1と同じものです。等々、あります。
最近割とよく耳にされると思いますが、UTF-8もエンコーディングの一つです。
日本語だけでも幾つもあります。いちいち読みません。
ちなみにCP932は、Windowsでのコードページという古くからの呼び名です。
Shift-JISとほぼ同じなのに、ほんの少しだけ違うというもので、これで問題を起こすこともままあります。

もっと細かく言えば、コード表とエンコーディングは別です。
同じコード表でも別のエンコーディングができたりするので、話は大変やっかいになります。
しかし、その辺の細部は今日の主題ではありません。

・スライド9「フォント」

「文字が化ける」という現象の大きな理由として、エンコーディングの違いが分からずに違うエンコーディングで見てしまうということのほかに、フォントで化けるというものがあります。
フォントというものも大抵の方は耳にされたことがあると思います。

昔、タイプライターがあった時代は、IBMの電動タイプライターは丸い玉の周りに活字が付いていて、これが一つのフォントのセットでした。
要するに、出せる文字の実際の形を全部ひっくるめてフォントとしていました。

MS Pゴシックは Windows で標準に付いているので、よくあると思います。
これを YouYuan という中国のフォントでやると、それほど大差はない形になります。
MSPゴシックで32666を出すと「羚」になって、YouYuanだとそれによく似た文字になります。
ところが、内部コードの70をMS Pゴシックで出すとFになりますが、それをWingdingsというフォン

トで出すとFではなく、指を指しているマークが出てきます。
なぜこんなことになっているかというと、かつて電動タイプライターでは、Wingdings はなかったのですが、ディングバットという玉に変えてFを押すとこの字が出てきたという経緯があり、本来ならあまり好ましくないのですが、こういうものが今でも残っているのです。
ところが、違う内部コード「9758」をMS Pゴシックで出すと、同じ指さしのマークが出てきます。
そういうものの重なりで、文字化けが起こることがあります。

・スライド10「いろいろなフォント」
先ほど、エンコーディングにどんなものがあるかをお見せしましたが、フォントの方も勘定しきれないぐらいあります。
普通にパソコンを買った状態で幾つぐらいフォントが入っているかというと、100や200ですが、世の中に流通しているフォントは何千とあります。

有名なのは、タイプライター時代からあるCourierや、Windowsで割合と使われるLucida Sans。
Times New Roman（タイムズニューローマン）というタイムズの新聞系のもの。
平成明朝と呼ばれる、コンピュータ用にできたようなもの。
韓国系のGungsuhや中ゴシックBBBなど、いろいろな名前のありとあらゆるフォントがあり、それによって実際に出てくる形が変わります。

赤丸を付けたLucida SansとTimes New Romanの「g」に注目してください。
どう見ても形が違いますよね。
Lucida Sansは9に似ているし、Times New Romanは8に似ている。
でも、同じ小文字のgだということは、少なくともアルファベットを知っている人なら誰も不思議に思ったことはないのではないかと思います。不思議に思ったとすれば、習い始めですかね。

似たようなことがここでも起きています。
情報学の「情」が、Gungsuhだと「青」の字の下の部分が「円」のように書かれています。
中ゴシックBBBだと、下に「月」がある「青」になっています。
でも、これは同じ字だと大抵の人は思います。

もう一つ。
HGS教科書体とHG明朝Bの「さいたま」の「さ」という平仮名は違うと思いますか。
HG明朝Bの「さ」は全部つながっている。HGS教科書体の「さ」は離れている。
形としては違うけれども、これを違う字だと言う人は、普通はいません。
しかし、これを出したのにはエピソードがあります。
さいたま市ができて、平仮名で書くことになったとき、さいたま市の「さ」はつなげて書くのが正しいのか、離して書くのが正しいのかと、役所に質問が来たのです。
伝聞ですが、事実のようです。

・スライド 11「文字とグリフ」

こういうことを説明するのに、コンピュータの周りで主に使われる「グリフ」という言葉を使います。
文字とグリフは別です。グリフとは字の形で、字体と言ってもいいです。

文字 103 番は「LATIN SMALL LETTER G（ラテン文字小文字の g）」という名前をしています。
その文字に対して、9 に近い形の「g」のグリフも、8 に近い形の「g」のグリフも対応します。

逆に、スライドの一番下の丸いグリフは、「LATIN CAPITAL LETTER O（オー）」の大文字の O（オー）と、DIGIT ZERO では数字の 0（ゼロ）の両方に対応しています。
昔、タイプライターの時代は、数字の 1（いち）の代わりに L の小文字、数字の 0（ゼロ）の代わりに O（オー）の大文字を打つのは、ごく普通のやり方でした。

このように、グリフとは形で、文字は形ではなく、それを抽象化したある一つの概念だと説明するために、こういう図を使います。

・スライド 12「いくつかの用語」

ここから先はコンピュータから離れて、文字自体の話に入ります。
そのために少し、言語学の用語を解説します。
と言っても、私は決して言語学者ではありませんので、必要な部分だけお話しします。

「スクリプト」という概念があります。
これは例で言う方が早いので、「ラテンアルファベット」がその一つです。
「キリル文字」はロシア語、セルビア語など、スラブ系の言葉で多く使われる、R のひっくり返ったような字があるものですが、あれも一つのスクリプトです。
右から書いていく「アラビア文字」も、一つのスクリプトです。
漢字もハングルもかなも、一つのスクリプトです。
若干問題になるのが、片仮名と平仮名は別のスクリプトか、同じスクリプトかということです。
これはなかなか難しくて、結論の出ない話になります。
そのような関係はほかの言語でもありまして、「半独立のスクリプト」という言い方をすることもあります。以上、スクリプトという言葉はこういうものです。

次は全く別の話、「話し言葉」と「書き言葉」についてです。
英語では Spoken Language と Written Language という言い方をします。
私がこうやってしゃべっている日本語と、書かれる日本語とは全く同じではないということは、感覚的に簡単に理解されると思います。
最近では、話し言葉と書き言葉のほかに、手話（Sign Language）なども独立の言葉の種類として認められています。これが用語の 2 番目です。

用語の 3 番目は「書法」です。
この言い方は、言語学で普通に使われている用語ではありません。
日本語で「書法」というものは、一番普通には、書道の世界で「書き方」のことをいいます。
ただ、ここでは正書法の書法というような言い方で、書き言葉を表現するための規則の集合です。
英語で Writing System という言葉がありますが、これはそれよりもう少し狭く、普通は書き言葉を表現するための文字体系をいいますが、それをもう少し広げて、「書法」という言葉を使わせていただきます。
これはこの講義限りだと思ってください。

そういう言い方からすると、書き言葉としての日本語は、漢字・かなの二つの別のスクリプトを交ぜて使うのが日常的な書法になっています。これは現在、世界のメジャーな言語のうちではただ一つです。
マイナーな言語を入れても、現在では恐らく日本語しかありません。
その意味で、一番難しいということは確かだと思います。ただし、話し言葉としての日本語が特に難しいという点はないですね。ほかの言語と同じぐらいのものです。

脱線しますが、最近の日本語は、漢字とかなだけでなく、ラテンアルファベットもごく普通に使いますので、それから言うと現在では三つのスクリプトを混用しているのかもしれません。

・スライド 13「漢字とは？」
さて、漢字というものはスクリプトなのですが、もともとは中国語用のスクリプトです。
これは誰だってご存じですよね。
1 文字が中国語の一単音節語、「チイ」といった一音節で表される語に対応しています。
そのために、表意文字（ideograph）ではなくて、表語文字（logograph）、語を表す文字という言い方をする方が、最近は学問的には主流かと思います。
考え方一つの問題ではあるのですが、字に意味があるのではありません。
字が中国語の単語と対応していて、その単語に意味があるのです。
これが常に 1 対 1 対 1 に対応していれば問題はないのですが、時々、字と語の関係がふらふらと離れてしまったり、離れることはあまりありませんが、二つがくっついてしまったりすることが起こります。
それが話をややこしくするもとの一つです。

一方、日本語の中での漢字は、これはもちろん書き言葉としての日本語ですが、「当時」がいつかは置いておいて、「当時の中国語の発音を日本語の音韻構造に移したもの」です。
舌をひっくり返して裏に付けた音を聞いても、日本語では書けないので、日本語で聞くとそれは「チエン」と書くことになってしまいます。
そういう変化を遂げたものが「音」です。
それから、その字が組み合わさっていた単語の意味の日本語訳が「訓」として使われるようになりました。
この二つを混用しながら表記します。

これをいつごろ始めたかは定かではないのですが、少なくとも万葉集の時代にはある程度形が整ってきていました。
その中で、中国語の単音節語にもともと対応しない「国字」があり、「畑」や「畠」がそうです。
そういうことをしながら、書き言葉としての日本語を千何百年か発展させてきたわけです。

・スライド 14「『漢字コード』の歴史」

一方、コンピュータで文字を扱うために文字コードが必要になりました。
コンピュータができたのが、1946 年の ENIAC が最初など、幾つか説がありますが、せいぜい 1900 年代半ばの話です。
それらが文字を扱えるかというと、50 年代のすぐには、最初は大文字だけでしたが、アルファベットを扱えるようになりました。
日本語が扱えるようになったのがいつごろかということは、覚えていらっしゃる方がおいでとは思いますが、1970 代半ばぐらいに急速に発展し、80 年代に入るとかなり普通に使えるようになってきたというぐらいのものです。

JIS C 6226 が、JIS として最初にできた漢字コードです。
「情報交換用漢字符号系」というぐらいですから、文字コードではなく、漢字コードといっていました。
今でもその言葉はよく使います。
別に漢字だけがあるのではなくて、この中にはアルファベットもあれば、ついでにロシア語やギリシャ語のアルファベット、平仮名や片仮名もあって、そして漢字もあるというものが 1978 年にできました。
その前は、なかったわけではありませんが、各社別々の文字コードを使っていたのです。

少し脱線しますが、私がコンピュータで漢字などを扱い始めたのは国文学研究資料館で、古い資料の目録をコンピュータで全部作ることになったのです。
最初にやったときはまだ JIS コードができていなくて、入力する会社のコードと、プリントする会社のコードの二つだけではなく、もう一つ、校正用のゲラを出すところのコード、三つが全部別々だったのです。
お互いコード変換して、校正用のゲラを出し、その校正用のゲラに赤を入れて、戻して、返ってくるとまた同じところが間違っている。
また赤を入れて直すということをやって、3 回目ぐらいに気が付いたのは、コード変換表が間違っていたということでした。
そういうことが、78 年に JIS コードができることでようやくなくなりました。

1983 年にそれが改訂されて、このときにいわゆる旧字と新字を入れ替えました。
当用漢字表と常用漢字表の入れ替えに伴って行ったのですが、これが議論を呼びまして、そこから漢字コードの話がわあわあと大騒ぎになってきました。
パソコンが使われるようになってきたのも、ちょうどこのころからなのです。
いろいろな人がコンピュータで漢字を扱うようになってきたころ、間の悪いことに当用漢字表から常用

漢字表への切り替えが行われ、ちょうどそのときこの改訂をして、文字をひっくり返してしまった。
それが議論の始まりだったような気がします。
この 1983 年の版は 87 年に JIS X 0208 と名前だけ変わったので、いまだにこの名前が普通に使われています。

1990 年に JIS X 0212 という補助漢字符号ができました。
そこで 5801 字を足したのですが、私はこれを決めるときの委員をやっていて、たった一人だけ反対したのですが、賛成多数で通ってしまいました。
実際、この漢字コード表は、ほとんど使われることはなく終わりました。

1993 年に ISO で、ISO/IEC 10646 UCS ができました。
そこでは 2 万 1140 の漢字が入っていました。
当然、最初の JIS X 0208、0212 はすべて含まれています。

次に 2000 年に、JIS X 0213、いわゆる第 3 水準と第 4 水準ができています。

・スライド 15「『漢字コード』の歴史」

一方、ISO では、UCS というコードを、漢字の部分は特にどんどん増えていき、2000 年には 6583 の漢字を増やし、全部で 2 万 8000 字、3 万足らずになりました。
2003 年になると幾つ増やしたかよく分からなくなったので、全部で 7 万 1226 の漢字。
2011 年、今年の春に正式に決まったものでは、7 万 5616 という漢字の数が、一応定義されています。
このあたりの文字コード関係は、京都大学の安岡孝一という先生が『文字符号の歴史―欧米と日本編』を共立出版から出しており、その本に非常に詳しく書いてあります。

・スライド 16「問題点：何が一つの『漢字』か？」

さて、勘定していったら七万五千幾つになると言いましたが、漢字の文字コードをめぐる議論の中でどうしても決着しないのが「何が一つの漢字なのか」という問題です。

例として、野原の「野」と、「埜」と書くもの。
宮澤の「澤」も、今の常用漢字表の字「沢」と、旧漢字「澤」と、中国の簡化字があります。
それから、上から 4 番目の字は両方「姫」に見えると思いますが、左が新字体、右が旧字体という言い方をすることもあります。
それから、文字のなべぶたの下が「口」になっている「高」と、はしごのようになっている「髙」。
われわれはこれを普通、「口高」「はしご高」と俗称しています。
江戸の「戸」も、上の「一」がくっついているものと、点のものと、離れているものがあります。
これらが、ISO 10646 UCS ではすべて別の文字で、別のコードを割り当てられています。

・スライド 17「問題点　続」

一方、葛飾の「葛」という字は、中が片仮名の「ヒ」のようになっているものと、「人」に特殊なカギが付いたものになっているものがあります。
「与」の新字体は「与」ですが、横線が突き通っておらず、少し斜めになって止まっているものもあります。
次に出した例が、「芸」が三つ書いてあるように思われるでしょうが、「藝」の常用字体である「芸」、もともと「ウン」という字の「芸」、また、中国で使う、くさかんむりに「雲」という字の簡体字です。
これらはすべて、ISO 10646 UCS では同一コードしか持っていません。
一体それがどこでどう決まっていて、どう分けられるのかということが、いつでも問題になっています。

・スライド 18「文字学」

そういう問題に対して、文字学という分野があります。
この際は漢字の話ですから、漢字関係の文字学です。
中国でずっと昔からありまして、漢の時代に既に「設文解字」という最初の辞書のようなものができて以来、漢字の字典は面々と続いています。
特に「諸橋大漢和」と呼ばれるものは戦前から始まって戦後に刊行されましたが、一つの金字塔という言われ方をしています。
最近では白川静という先生が『字統』という字源学的研究を取り入れた漢字字典を出しています。
このように、文字学はそれなりに非常に発達した学問です。
これらできちんとすれば、文字コードの混乱は起きないのではないかと普通は考えるわけです。

・スライド 19「野埜」

「野」を『康熙字典』が何と説明しているか。
『康熙字典』は 18 世紀の清の時代にできた、康熙帝という皇帝が編さんさせたもので、現在、世界的に漢字では最も権威のある字典ということになっています。
それがどう説明しているかというと、「野」の古文として「埜」があるとされています。
次に『大漢和辞典』、「諸橋大漢和」では別の文字として立ててあり、見出しが別にあるわけです。
「埜」は右側の「野」という字になりました。
それから白川先生の『字通』という本では、一つの見出し「野」の下に、「埜」などが別の字形として挙げられています。
『漢語大字典』という中国の今一番大きな辞典は、細かいところは分かりませんが、大体 1980 年代に出されたものだと思います。
やはり「大漢和」と同じように、この二つの「野」「埜」は別々の見出しで立っています。
「埜」は、何と言っていましたか、古体（こたい）というような言い方で説明しています。

では、これは一つの字なのか、別の字なのか。

・スライド 20「姫姬」

この辺をゆっくりやっていますと、とても 1 時間では終わらない。

1 文字だけでも、これを作るだけで大変な時間をかけているのです。

「姫」の字が二つありまして、おんなへんの横に「臣」を書く方と、中の「口」の左の線が離れて、複雑になっているものがあります。
実は『康熙字典』の「姬」の中も複雑になっていて、それらが両方あると言いましたが、『大漢和辞典』で見ると、われわれが普通に「姫（ひめ）」と思っているのは「シン」という字で、「つつしむ」という意味がある。
複雑になっている方が「キ」という字で、「姫」という意味だとされています。
ただし、「姫」の「参考」に、もともとこれは別だけれども、今は「姬」の略字としてこちらを用いると書いてあります。
この辺は「諸橋大漢和」の辛いところなのですが、諸橋さんが『大漢和辞典』の編集を始めたのは戦前で、戦後になって刊行したときに当用漢字が出てきて、わあわあと混乱した状態になりました。
その途中で編集をやっていたのですから、これは大変だったろうと思います。
その結果、説明がこういうことになったのです。

白川先生の『字統』によると、複雑になっている旧体は「姬」の別字形としているのですが、白川先生の『字統』『字通』も一応、常用漢字表を認めており、見出しはすべてその形に従っているので、このようになります。
当然のことながら、もともとの形からすると、中に一本入って複雑になっている字の方が、本来あるべき字形だと考えられておられたのは確かだと思います。

『康熙字典』では、この二つは見出しとして全く別の字です。
中国の『漢語大辞典』では、最初から別の字です。

・スライド 21「姫姬　用例」
しかし、実際に「ひめ」という字が書かれているいろいろなところを調べると、必ずしも今の常用字体が間違いとは言えません。

『大書源』は、書道のために古くからいろいろな人の書いた、あるいは鐫刻（せんこく）された漢字を非常に多く集めた本です。
その中でずっと見ていくと、中にはどう見ても今の新字体と同じような書き方をした「姫」があります。
また、今の新字体の「臣」と「女」の間に、一本、縦に棒が入ったものもあります。
ほかにも、今の新字体と同じように書いているものがあります。
一方、旧字体に書こうとしたと見られるものもあります。
また、光明皇后が書を書いた本から出てきた字は、このような新字体というか、新の形で書いてあります。こういうことが普通にあったのです。
ここだけでも 3 種類ぐらいの形が入っています。

・スライド 22「高髙」

一方、「高」という字の「はしご高」と「口高」に関しては、『康熙字典』では「口高」しか認めていません。『字統』もそのとおりです。

ただし、書いてある字の形は「はしご」に見えるものもたくさんあります。

『大漢和辞典』では別の見出しで立てており、はしご高は口高の俗字だと言っています。

『漢和大辞典』では、別々に字を立てています。

・スライド 23「高髙　用例」

実際にはどう書かれていたか。目を皿にしても見て仕方がないので、雰囲気だけ。

わざと楷書風から、行書風、草書風、隷書、篆書風といろいろなものを出していますが、人が書くときはさまざまなのです。実際に非常に多くの形で書かれています。

どちらかというと、はしごの方が優勢ですね。

実は、「宮澤」の「宮」も、口と口の間にチョンと入って「呂」になっていますが、「ノ」がなく口が 2 階建てになっている形があって、昔の書かれたものを調べると、そちらの方が優勢です。

ただし、『康熙字典』では「ノ」が入っています。

だから後で、そちらの方が正しいということになったということです。

・スライド 24「葛葛」

次に「カツ、くず、かずら」の字は、『字通』では当然一つだけです。

ほかの字典でも、すべて中の部分が「人、カギ」の形で見出しが出ています。

ただし、それが実際にどう書かれたかというと、かなりのものは「ヒ」で書いてあります。

また、「エ」のようになっているものもありますし、『康熙字典』風の字体になっているものは決して多数派ではないですね。本当に書かれたものを全部調べるわけにはいきませんので、難しいのですが。

少なくとも一つぐらいは、江戸時代にどう書いていたかを調べました。

片仮名の「ヒ」で「葛飾戴斗」（北斎の別名）と書いてありました。

ところが、江戸時代の本には広告が出るのですね。

葛飾戴斗のこういう本が出ますよと書いたときには「葛」で書いていたのですが、実際にその版本が出たときには、「人、カギ」の字体で書いてありました。

・スライド 25「文字学は」

文字学でこういうことをやっていくと、きちんと一貫した文字コードが決められるかというと、人や字典によって言うことがさまざまなのです。

白川先生は白川先生で、きちんと一貫した答えがあると思います。

「大漢和」になると、少し怪しくなります。

これは、編さん途中で当用漢字表ができてしまったという事情もあるからだと思います。

ほかの字典も入れますと、どんどん分からなくなります。

私は、心情的には白川先生風に「字は字体が違っても全部同じだ」と言いたいところですが、そういうことを言うと、「野」と「埜」は同じ文字コードだ、グリフが違うだけだということになります。
それが実用的かと言われると「うーん」と首をかしげますし、恐らく、ある程度分かった人を 10〜20 人集めれば、「別にした方が、問題がないだろう」という答えが過半数を占めると、ほぼ確言できます。

それから、辞書に示されていない字体が実際にはたくさんあるわけです。
これは昔の書や版本などを調べるとごろごろ出てきます。
「百寿」といって、「寿」の字を 100 種類の字体で書いた本もありますから。

・スライド 26「異体字」
こういうものを総称して異体字といいますが、先ほどの用語に戻って言うと、「異なるグリフ」なのです。
昔から「古文」「古字」「俗字」「異体」など、いろいろな言い方で言い習わされてきたものです。

もう一つ、はしご高と口高のように、隷書体から楷書に写すときに筆画をどう決めるかの違いによって生まれてきた異なるグリフに対しても、異体字という言葉を使います。
実はこれが非常に多いのですけれども。

3 番目が、政治的と言ってもいいのでしょうが、漢字簡素化です。
日本での当用漢字や常用漢字、中国での簡体字。
昔、則天武后が則天文字を作ったそうですが、そのように書法を変更することによって、新しいグリフが生まれてしまったものも異体字の仲間です。

・スライド 27「字体の『包摂』」
それで議論をしていたら、恐らく 20 世紀中には漢字コードは生まれなかっただろうと思います。
漢字コードを作る人々は、仕方がないから「字体の包摂」という概念を用いてこれを何とかしようとしました。
「包摂」という言葉をここで使ったのは豊島正之先生で、JIS X 0208 の 1997 年版から「包摂規準」を持っています。
それから UCS（ISO）の方では、unification rule という名前で似たようなものを決めています。

・スライド 28「Unification ルール（Unify の例示）」
その中身は、要するに大きな表があって、この字とこの字は同じにします、この字とこの字は別だと思いますというものの例示なのです。

・スライド 29「Unification ルール（Separate の例示）」
unification rule で、このスライドに示した字は違うということにしてあります。

・スライド 30「JIS X 0208-1997 の包摂規準（部分）」
JIS X 0208 の包摂規準は、何十ページかにわたってこのような表がずっと続いています。

・スライド 31「包摂の問題点」
そのような規準を決めてやってきたのですが、包摂規準を決めるより前に、既に漢字コードはあったのです。
実ははしご高と口高は、台湾のある漢字コードで、別の文字として決められていました。
それらを取り込んで UCS を作ったとき、もと別であったものは一緒にしてはならないという規則を決めてしまったために、別の文字コードということになりました。

これと同じようなことは、JIS X 0208 の包摂規準でも非常に苦しい説明をしていますが、もう決まってしまっていたわけです。
決まっていた後から包摂規準を作ったので、どうしても例外にならざるを得ないものがあることが問題点です。

2 番目の問題点は、そうは言っても形だけから文字を決めるのは、やはり本来あるべき姿ではありません。
形の包摂規準だけで、文字を全部識別することは無理です。
本来、文字とは、形とは独立にあったものです。

3 番目の問題は、「万人が納得するものは決してできることはないだろう」ということです。
これは言語のほとんどの場面で言えます。
「言う」という言葉を「いう」「ゆう」のどちらが正しい書き方かと聞かれて、万人が同じ答えを出すことはまずあり得ない、そういうたぐいのものです。

・スライド 32「結論として　1」
さて、結論です。文字は形ではないのです。
グリフは形ですが、文字は、そのグリフを幾つか含み得る抽象的な構造物です。
白川先生は、「康熙字典の［一］部には丁、丐、七・・・を録する。・・・字はその構造的な原理から離れ、その構造的な意味も捨てられて、ただ筆画の形式によって分離配列されている。そこにあるものは既に文字ではなく、文字の形である。意味を失っている記号である」
と言っています。

しかし、実際のところ、『康熙字典』の時代で、既に文字は形になりかかっているのです。
単に記号になりかかっている。
楷書の形にまとめていく段階で、どんどん記号としての使われ方をされるようになってきています。

・スライド 33「結論として　2」

また、社会的合意が弱い。「弱い」とは、問題になったときに圧倒的多数で結論が出るということが少ないということです。

では、どうしたらいいのかというと、ここの技術的解決法は、ないことはありません。
現在の IS0 10646 の包摂で不都合な場合は、アプリケーションソフト側で正規化を行うか、一つの文字の中で異なるグリフを区別するためのフォントの切り替えを行うことによって、何とか解決というか、逃げることができます。ただし、この技術を解説し始めると、これはこれで 1 時間では終わらない話になるので、今日はここまで。

・スライド 34「結論として　3」
3 番目、私の一番言いたいことです。
文字の形の本当に小さな違いを問題にして、どちらが正しいのだ、こう書かなければいけないということを一生懸命、どんどん細かくしていきたがる社会的風潮を、私は大変に憂えております。
それは取りも直さず、人々が漢字をよく分からなくなったからです。
漢字をもっと知っていれば、こう書かれても、ああ書かれても、「これは同じ字なんだ」と皆が理解して、特に問題は起きなかったでしょう。
それがよく分からなくなったから、さいたまの「さ」を区別したがるなどということになってきました。

包摂規準を決めた豊島先生に対して、林大先生という JIS を決めたときの偉い先生が、
「これを作ったときには、包摂は社会的常識だと思っていた。そんなことが問題になるとはほとんど思っていなかった」と言っています。

ところが、コンピュータができて、皆でやり始めた 80 年代ぐらいになると、そういうことを非常に問題にするようになってしまった。私はやはり嘆かわしいと思っています。そういうことで頭を使うのは、知的資源の無駄遣いです。

ということを最後の結論にしまして、以上で終わりです。
ありがとうございました（拍手）。