人文科学とコンピュータ 28－4
（1995. 11. 25）

インターネットにおける学術漢字の符号化に関する基礎的研究

斎藤秀紀・柳沢好昭・横山詔一

国立国語研究所
情報資料研究部
日本語教育センター

技術教育を効率よく行うためには、日本語資料に対する提供手段の開発と分野別で使用される漢字の特長を知ることが重要である。本稿では、最初に構造化４バイトコードが大規模の漢字データベースと専門分野別文字集合を一つの枠組みで符号化できることを示した。次に、学術漢字の教材化に適用可能な出現頻度・親密度データベースを紹介した。最後に、海外の日本語学習者が情報を収集し、その体系を発見することに有用であり、かつ、学習者自身が収集した情報から簡便に教材作成が行えるプログラムとリソースのサンプルについて述べた。

Basic Research on the Encoding of Kanji for the Transmission of Technical Information Via Internet

Hidenori Saito, Yoshiaki Yanagisawa, Shoichi Yokoyama

The National Language Research Institute
Department of Data Orientation
Center for Teaching of Japanese as a Second Language

In order to improve the efficiency of education in technical subjects, it is essential to develop the means to transmit technical information in the Japanese language, and to examine the characteristics of Kanji used specialized fields. In this study, we show that it is possible to use a structured four-byte code to encode both a large scale Kanji database as well as sets of characters used in specialized fields, within a single framework. Next, we describe a database which includes indices of frequency and familiarity of characters which could be used in adapting literature using specialized Kanji for educational purposes. Finally, we describe a software system for allowing people studying Japanese overseas to accumulate and organize their personal observations from a variety of media, looking for rules and systematic regularities which connect them.

Ⅰ　学術漢字の符号化の方法

１．１　はじめに

本研究では、最初にＪＩＳ Ｘ０２０８およびＩＳＯ／ＩＥＣ１０６４６の問題点を示し、大漢和辞典の検字番号を基本に漢字符号に求める機能を３ 種の構造をもつ４バイトコードで表現できることを示した。次に、４バイトコード（整数部３バイトと小数部１バイト）の小数部に異体字と中国・台湾・日本・韓国の情報交換用文字を配当する１字体１符号化法が、文字集合の全体から部分集合を抽出し、２次符号化する方法に拡張できることを述べた。さらに、文字集合の部分は、専門分野別に漢字を符号化できることを示した。

情報交換用漢字符号ＪＩＳ Ｃ６２２６ は、（１９８７年ＪＩＳ Ｘ０２０８ に名称変更）１９７８年に日本工業規格として制定された。現在ＪＩＳ Ｘ０２０８は、利用方法の多様化とともに、多国語表現や古典・漢籍などにも使用されている。しかし、大規模の漢和辞書を始め古典・漢籍を電子化するためには、文字種の不足が指摘されていた。また、１９８３年の規格改正では、水準間で２２組みの字形入れ替えと２９４字の字形変更があり、新旧規格の間で互換性を崩した。問題には、符号間への文字の追加機能や変更情報の局所化など外界の変化に対する調整機能や、文字と符号に対する規範性に対する考え方の曖昧さ、異体字の整理に関するものがある(斎藤:1985, 1993)。

一方、国際標準化機構は、世界共通の情報交換用符号の利用を目的に１９９３年にＩＳＯ／ＩＥＣ１０６４６－１を発表した。しかし、ＩＳＯ／ＩＥＣ１０６４６－１は、中国・台湾・日本・韓国で使用されている国内規格を２バイトコードの２０、９９２の枠内におさめようとしたため新たな問題が指摘されている。問題は、各国の文字政策の変更とＩＳＯ／ＩＥＣ１０６４６－１ への反映方法や、漢字の読み・画数・部首などの属性情報の規定に関するもの、既存の漢字符号からの移行や、各国語の混在と個別処理を指定に係わるもの、符号化規則に対する「ソースコード分離」や「字形の統合」の併用に係わるものがある(斎藤:1994a, 1994b)。

１．２　４ バイトコードの構造化

４バイトコードの符号化領域は、ＪＩＳ Ｘ０２０８（Ｇ０領域）や拡張ＵＮＩＸコード（Ｇ１領域）との併用を考慮し、Ｇ３領域を使用する（図１）。Ｇ０からＧ３の各符号の識別は、２の８ビット目を使用した。１０進数で表現された検字番号の内部符号への変換は、９４進数と１６進数の二重変換を行い（以下１６進化９４進数変換と呼称）１６進数‘２１’から‘７Ｅ’に調整した。符号化できる字数は、整数部８３０、５８４字と小数部９４字である（式１、２）。４バイトコードの構造は、整数部が（１）漢字辞書やコードブックに付けた１０進数５桁の検字番号を３バイトで表す構造（図２－１）、（２）既存の２ バイト系漢字符号を切り替え符号で統轄する構造（図２－２)、（３） 各桁の２の８ビット目が‘０１’である２バイトコードを２個使い内部符号とする構造 （図２－３）、（４）２バイトコードと４バイトコードの各桁２の８ビット目を識別用とする４種である。小数部は、整数部に１バイトを付加した構造を基本に、（５）異体字と各国語の情報交換用文字を配当する構造である （図２－Ａ）。

整数部＝ＨＥＸ（検字番号ＭＯＤ９４）＋重み‘２１’　　（式１）

小数部＝ＨＥＸ（小数部挿入位置番号）＋重み‘２１’　　（式２）

ＨＥＸ:１０進数値を１６進数に変換する関数

ＭＯＤ:１０進数表現された検字番号の剰余を求める関数

検字番号:１０進数５桁で初期値が‘０’の数値

G0 I－00	G3 I－01	表外字コード領域 G3
G2 I－10	G1 I－11	内字コード領域 G0、G1、G2

図1　2バイトコードと4 バイトコードの符号化領域

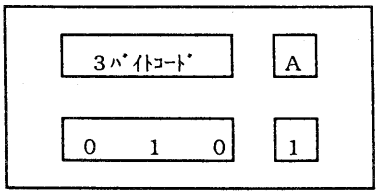

図2－1
検字番号を4ﾊﾞｲﾄｺｰﾄﾞ 化した構造
符号の識別ﾋﾞｯﾄ 列

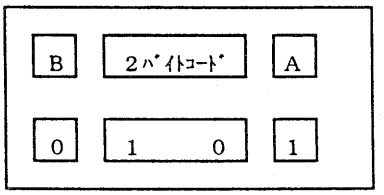

図2－2
ＪＩＳ Ｘ０２０８を基本とする符号の構造
Ａ：文字追加用符号
Ｂ：各領域の識別符号

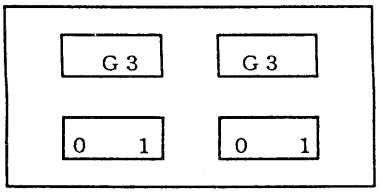

図2－3
Ｇ３を基本とする内部符号の構造

図2　4バイトコードの構造

1．3　1字体1符号化法

漢字符号の1字体1符号は、4バイトコードの整数部3バイトに見出しに相当する字体（特定の文字で代用）を対応させ、中国・台湾・日本・韓国の情報交換用文字集合や異体字を小数部1バイトに配当する方法をいう(斎藤:1994a)。この符号化法は、変更の多い情報を見出しから分離し、見出しと漢字符号の関係を安定させる効果がある。また、旧規格の漢字符号で作成したデータの変更情報の履歴を小数部に累積させることによって継続使用を可能にする。4バイトコード小数部への文字配当は、符号化の対象になる文字集合を部首順に配列したのち小数部‘1’から‘4’に中国・台湾・日本・韓国語の情報交換用文字を登録する（図3）。‘5’から‘94’には、新しい文字の追加機能を基本に見出しに対応する異体字の全リストを登録する。

見出し	情報交換用文字				異体字登録領域									
0	1	2	3	4	5	6	7	8	9	10	11	12	13	----- 94
剣	G	T	J	K	剣	劍	劒			劔	劍	劒	釼	

図3　小数部による文字指定の方法（中国：G，台湾：T，日本：J，韓国：K）

１．４　基本漢字の選定と符号化の方法

４バイトコードで表現できる文字集合は、多様な文字集合を符号化できるが実務で使用するためには柔軟性に欠ける。文字集合や符号は、目的にあわせ自由に設定できることが重要である(斎藤:1995)。分野別や専門別に文字集合を規定することは、専門的な読みや漢字を一般利用者に理解させる必要がなくなり、利用者の負担軽減と外国人に対する専門分野別の教育への利用が期待できる。利用者規定の方法は、図２－１に対応する文字集合を規範とし、図２－２の構造の枠組みのなかで２バイト符号を与える。異なる符号間での情報交換は、規範となる４バイトコード対応の文字集合を中間符号に使用する。

現在、情報科学で使用している日本語から基本漢字を選定する作業の一貫として、日本語・中国語・英語で使用される用語を入力している。この資料に対して、以下に述べる分析を試み、日本語学習リソースの開発を目指している。

Ⅱ　学術漢字の認知的特性の解析に向けて

認知心理学では、語彙の習得において学習材料の出現頻度や親密度（familiarity）が学習効率に大きな影響を及ぼすことが知られている。親密度とは、学習者が当該の言語刺激にどの程度見慣れているかを示す心理的指標を言い、評定尺度法で測定される。

日本語学習者のための語彙学習教材を開発するには、語彙の出現頻度や親密度の資料が重要な役割を果たすと考えられる。以下、日本語学習の分野に適用が期待できそうな出現頻度・親密度データベースを紹介し、学術漢字の教材化にどのように応用できるか簡単に述べる。

＜出現頻度データベース＞

漢字および漢字単語の出現頻度については、国立国語研究所の語彙調査が諸学界から高い評価を受けてきた。しかし、それらは約３０年以上も前に実施されたものであり、時代的に現状にあっていないのではないかという声を耳にするようになってきた。国立国語研究所の語彙調査データは１９９０年代にも果たして通用するのであろうか。

この疑問に答えるため、野崎・横山・磯本(1995)、野崎・磯本・横山(1995)は１９９３年の朝日新聞記事ＣＤ－ＲＯＭテキスト・データベースを利用して漢字および漢字単語の出現頻度を調査中である。現在、１９９３年の朝日新聞記事ＣＤ－ＲＯＭに格納された約１１万件の記事に登場する漢字１文字の出現頻度について集計を終えた。次の段階では、国立国語研究所が実施した１９６６年の新聞語彙調査データとの相関を分析し、漢字の出現頻度に時代差がどのような影響をもたらしているのか明らかにしていく。

＜親密度データベース＞

親密度については、これまで大規模なデータベースの整備には手が付けられていなかった。しかし、Amano, Kondo & Kakehi(1995)により約６万の見出し語を有するデータベースが構築されつつあり、状況は好転した。このデータベースは、文字データベース、単語データベース、音声ファイルデータベースの３要素から成り、それらが相互に参照可能である。文字データベースにはＪＩＳ Ｘ０２０８に規定された約６８００文字に対する親密度が納められている。単語データベースには単語を文字で被験者に呈示したときの親密度と、音声で呈示したときの親密度のデータのほか、アクセント、音韻表記、文字表記などの情報が約６万語のそれぞれについて含まれている。音声ファイルデータベースには、単語データベースの見出し語に対応する音声情報が格納されている。

＜学術漢字の教材化に向けて＞

以上で述べた出現頻度データベースと親密度データベースを利用すれば、情報科学分野における学術漢字の認知心理学的特性をかなり的確に把握することが可能となる。その知見に基づいて教材開発に着手すれば、合理的な日本語教育教材を手にすることができると期待される。

Ⅲ　日本語学習リソース開発

従来の学習観では、外国語習得は学習コースに参加しクラスワークを行うもの、教師によるコースデザインがあり、それに沿って学習を進めるのが一般的である。しかし、学習により新たな知識や技能を獲得するという学習効果が海外の様々な学習者すべてに同じように、どのような学習においても、いつでも現れるわけではない。学習者の個人差やニーズに対処するには、できる限り多くの個人差の次元（学習活動：学習者の認知的・情意的前提能力、学習課題、授業活動、到達水準、学習速度、情意的成果等、個人内要因：知能、認知スタイル、性格、情意的認知、動機付け、原因帰属、効力感、統制感、やる気、関心度、集中度、適性、交友、生活意識、情報交換網、自己理解等）について考えることが前提となる。

一方、現地国の日本語指導者、あるいは仕事や研究や将来のために学習する一般成人は、自分の学習や研修について検討し、全体計画を立て自律的に学習や研修を行う能力を備える。法則性や体系化の習得は、一方的に伝えるより学習者側が発見する方が定着度と情意面において効果的であり、いろいろな情報に接し、検索、複眼的に整理、分析、組み合わせて新しい発見を生み出すと考える。

学習を、（1）空間的知覚能力、記憶力、精神速度、推理能力、運動能力等の基本的能力からとらえる活動、（2）収集した言語行動情報に存在する規則や体系の結論を学習者自身が発見する過程に参加する活動、（3）日常の情報の能動的処理能力を尊重し、複数の人間が協働して解答を創り出す活動、を前提に、学習支援システムをコンサルトと人的・物的・社会的リソースと分け、視聴覚的アピール度、付加情報のインデックス化、開架式、集積・再利用と自己開発の簡便性を持つリソースと、その交流について試行する。その際、コンピュータ・ソフトは、操作の簡便性、視覚的アピールの高さ、ハードへの依存度の低さ、アクセスの容易さを要件とする。

本研究では、各種のメディアを合成・編集するツール、「Intellegent Pad」（開発者：北海道大学工学部田中譲教授、以下、ＩＰ）を用いてリソースを開発する。ＩＰは、すべての事物、絵やテキストや演算などのプログラムあるいはアプリケーションプログラムをパッドで表すものである。そのパッドは、それぞれ個別の機能を持ち、ＣＲＴ上で紙面のごとく表され、マウスで移動、複写、結合・貼り合わせ、拡大縮小や固有の機能を操作ができ、コンパイルなどの操作なしで機能合成や複合が行える。ＯＳやハードの垣根をも越える。このツールを使い、サンプル作成を行う。

作業１：日本語学習者の耳の上に小型ＣＣＤカメラを装着し、本人が希望する行動場面を録画し、そのデータをキャプチャーする。リンク・パッド、動画再生パッド、ビデオ／サウンド・パッド、開発ツール・パッドなどを使い、文字化資料、語彙や画像などのデータベースをリンクしたマルチメディア型素材を学習者自身が主体的に作成し、蓄積する。

作業２：学習者と指導者が作成物を視聴し、think-aloud による内観活動を行い、素材の追加データとする。

作業３：上記素材の送受信を行う。

【構成図】

録音録画データ	ﾋﾞﾃﾞｵ/ｻｳﾝﾄﾞ・ﾊﾞｯﾄﾞ	＊学習者自身が録画したもの（CVD-500 と CCD-MC1(SONY)）
文字化データ	ﾃｷｽﾄ・ﾊﾞｯﾄﾞ	＊学習者が主体的に書き起こしたもの
語彙データ	辞書ﾊﾞｯﾄﾞ	＊既存の語彙データベースの活用
文化情報データ	ｲﾒｰｼﾞ/ﾑｰﾋﾞｰ/CDﾊﾞｯﾄﾞ	＊非言語行動・文化事情等のデータ
付加情報データ	ﾃｰﾌﾞﾙ・ﾊﾞｯﾄﾞ	＊指導者等が補足的に付け加えた練習問題等
追加情報データ	ﾃｰﾌﾞﾙ・ﾊﾞｯﾄﾞ	＊think-aloud 等のデータ

ベース・パッド	
カーネル（Intellegent Pad ）	
ＯＳ （Windows3.1/ 漢字 Talk7.0/System7.0 ）	Intellegent Pad for Mac 1.0、漢字 Talk7.5
ハードウェア （IBM/ATorMac ）	PowerMac 8100/80AV(Apple)、1GB(HDD)、32MB(RAM)

参考文献

1）斎藤秀紀(1985)「漢字コードの拡張法に対する試案」『研究報告集６』国立国語研究所報告 83, pp. 57-103

2）斎藤秀紀(1993)「漢字コードのメタコード化の方法」『情報処理学会第４６回全国大会論文講演集(1), pp. 23－24』

3）斎藤秀紀(1994a)「１字体に１符号を対応させる漢字符号化の方法」『計量国語学会誌第 19 巻５号, pp. 223-233 』

4）斎藤秀紀(1994b)「大漢和辞典の検字番号に基づく構造化４バイトコードの提案」『情報処理学会論文誌 Vol. 35, No. ６, pp. 1119－1121』

5）斎藤秀紀(1995)「４バイトコード対応文字の部分文字集合に対する利用者規定の方法」『情報処理学会第５０回全国大会論文講演集（３）, pp. 201-202 』

6）財）札幌エレクトロニクスセンター（1994）「STEP NEWS vol. 83」

7）中野照海編著(1979)「教育学講座第６巻：教育工学」学習研究社

8）織田守矢ほか(1990)「学習環境の構築」コロナ社

9）今栄国晴編著(1993)「教育の情報化と認知科学：教育の方法と技術の革新」福村出版

10）アルフレッド・トマティス(1994)「人間はみな語学の天才である」アルク

11）後藤忠彦(1986)「教育とコンピュータ５：コンピュータと教育情報システム」東京書籍

引用文献

1）野崎浩成・横山詔一・磯本征雄(1995)「ＫＷＩＣ作成プログラムの開発」『教育システム情育報学会第２０回大会発表論文集, pp. 211-214』

2）野崎浩成・磯本征雄・横山詔一(1995)「日本語教育のための語彙調査」『日本教育工学会第１１回大会発表論文集』

3）Amano S., Kondo T., & Kakehi K. (1995)'Modality dependency of familiarity ratings of Japanese words' *Perception & Psychophysics*、57(5), pp. 598-603