CNS 漢字による部首画数データベース

川幡 太一
東京大学大学院理学系研究科
情報科学専攻

CNS 漢字は4万8千字の漢字に符号を定めており、また情報交換のための体系も整備されている。しかし、これらの漢字を簡単に扱う方法は整備されてこなかった。そのため、CNS 漢字に部首と画数のデータとそれを使って漢字を入力する方法を作成し、併せて異体字からの入力、および検索を行えるようにした。

1 はじめに

近年、コンピューターの発達に伴い、その応用範囲が広まってはいるものの、大規模かつ大量の漢字を編集・照合・交換するの一般的な方法はあまりない。台湾の漢字情報交換用符号規格である CNS 11643[5] は、約4万8千字の漢字に対して符号化を行っている。また、ISO 2022 [1] の方式で符号化を行っているため交換性も高く、ＥＣＭＡの文字セット集合 [2] にも登録されているため、一般的な文字セット切り替えで情報の交換が可能である。さらに 40x40 ドットのフォントも公布されていて、その入手も容易である。

公布されている漢字フォントの書体自体は台湾の書体となっていること、日本の国字にはほとんど対応していないなど、幾つかの制限はあるものの、中国古典に関係する資料の入力時における漢字の不足分を補うものとして有効に活用できると考えられる。

しかしながら、これを使って文字を簡単に入力・編集・検索などを行う環境は、今までは存在しなかった。そのため、CNS 漢字を容易に扱えるように、CNS 漢字に対して部首画数のデータベースとそれを利用して漢字入力を行う環境を整備し、また異体字による入力、検索を多言語エディタである Mule 上で簡潔に行うためのツールを作成し、またデータを一般に公開した。

2 CNS 漢字とその特徴

ここでは、台湾の漢字情報交換規格である CNS 11643 で符号化された漢字を CNS 漢字と呼ぶことにする。

CNS 11643 は 1992 年に改訂された際、約4万8千字の漢字に対して符号化を行った。これらは、ISO 2022 の基づく 94*94 のプレーン7面に漢字が定められている。このうち、最初の2面については、台湾の別の漢字符号規格である Big5 とほぼ同等である。

基本的に漢字はその使用頻度に基づいて、各面に割り当てられているが、最後の6面と7面に対してのみは、同じグループの漢字を両面に渡って画数の順番で配列している。

CNS 漢字には以下のような利点がある。

- 約4万8千字の漢字を扱うことができる。
- X-Window と Mule で使用できる。
- フォントが公布されている。
- 情報交換が比較的簡単に行え、特にインターネットでメールや WWW の文章の中に使うことが可能である。

CNS 漢字は UNIX 上で動作する X-Window システムと、多言語エディタ Mule の組み合わせで使用することが可能である。また、フォントも Mule の配布サイトから入手することが可能であり、これを使うことによって TEX 等でも印刷を行うことができる。Mule 上における、CNS 漢字の内部での処理などについては、[6] に詳しい。

また、メールなど、インターネット上での利用に関しては、CNS 漢字のテキストのインターネット上での交換方法が [8] で提案されているほか、Mule では日本語のテキスト情報交換方法でもある [7] の ISO 2022 の自然な拡張として、日本語にも混在して扱うことが可能である。

さらに、WWWに関しては、多言語対応ブラウザならば CNS 漢字を含む文書を読むことが可能である。一般のWWWブラウザの場合でも、例えばプロキシーサーバに Delegate [9] を使うことにより、CNS 漢字を含む HTML 文書を中継する際に CNS 漢字のみ相当の GIF 画像に置き換えることにより、読むことだけは可能にすることができる。

4万8千という数は必ずしもあらゆる漢字を包含しているわけではないが、通常の文献に現れる漢字に対しては十分である。また、フォントは日本の書体には合わないなどの問題もあるが、Mule では日本と中国の漢字は混在して扱えるので、必要の応じてこれらを混在することができる。

CNS を使う上での最大の問題は、その4万8千字全てにわたって適用可能な便利な入力ツールがないということである。そのため、CNS 漢字全てに対応する、漢字を入力するのに必要なデータベースと入力プログラムの整備を行うことにした。

3 CNS 漢字の部首画数データベースと入力・検索ツールの作成

4万8千字という膨大な漢字に対して、基本的な漢字の情報として、部首と画数によるデータを入力することにした。部首と画数は、漢和辞典の基本的な配列の方法でもあり、一般に読めない漢字の検字を行う際の最も基本的な情報である。

CNS 11643 の規格書には、規格書にある全ての漢字に対して、部首と画数が定義されている。そのため、規格書で定めてある各漢字の部首と画数に対して、文字認識を利用してデータ化し、これを使って入力するためのツールとして、Mule の「たまご」のメニュープログラムを利用して漢字入力システムを構築した。

部首と画数で分類を行った結果、全部で4万8千の漢字に対して、３８２７個のグループに分類できた。

漢字の中には正確に部首や画数を定めるのが困難なものがあることや、部首と画数の組合せによっては、候補が多いものがあるという問題点はあるものの、多くの場合において、必要な漢字の検索の範囲は大きく絞られ、必要な漢字をより容易に見つけられるようになった。

このデータを検討すると、これら３８２７個の部首画数グループのうち、候補が１００以上あるグループは５８あり、漢字の候補が最多のものは、くさかんむりの１３画で、漢字の候補数は２２０個になる。さらに、これら５８グループに含まれる漢字の総数は７５００字になる。候補が５０以上のものに属する漢字の総数は2万にもなる。

したがって、より速く簡単に漢字を見つけるためには、部首画数以上にさらに漢字をより細かく分類できるようにすることが望ましい。現在のところは4万8千字全てに対して与えられているデータは部首と画数しかないため、どのようなデータを補助的に与えるかは将来の検討課題である。

その一方、部首画数による以外の入力方法として、異体字などによる入力方法も用意した。CNS 漢字における異体字のデータベースとして、安岡氏による異体字データベース [3] などがある。これを利用して、入力中、または編集中の任意の漢字に対して、この漢字データベースへ参照し、適切な異体字が見付かるとそれに対して置き換えを行うことができるようにした。

この異体字関係は、漢字コードの種類の区別を超えてデータベース化されているため、JIS 漢字の補助として CNS 漢字を使用する場合において、対応する漢字が JIS にあるかどうかの確認を編集時に行うことが可能になる。

さらに、これらの異体字データベースを検索にも利用することにより、日本語・中国語のテキストを問わずに任意の漢字でＣＮＳ漢字とＪＩＳ漢字が混在した文章を検索できるようにした。

なお、この部首画数データベースの内容については、近日中に Mule への contrib として公開する予定である。

4 まとめ

最初に挙げた理由により、CNS 漢字にはその使用と情報交換において利点があり、漢文などの資料を入力する場合において必要な漢字がない場合、それを CNS 漢字で補うのは、一つの有効な方法であると言える。

しかし、これらの膨大な漢字に対して有効な漢字入力方法がないため、これまでは CNS 漢字を入力したりすることは困難であった。

そこで、CNS 全漢字に対して部首画数のデータベースとそれを利用して漢字を入力するシステムを作成した。また既存の異体字データベースと組み合わせることによって、異体字の漢字の検索と入力を容易にした。

将来の課題としては、より簡単に必要な漢字を多くの候補から見つけられるよう、他のデータを整備することなどが挙げられる。

5 謝辞

最後に、テスト公開中にこの部首画数データのチェックを行ってくれた方々に感謝いたします。

参考文献

[1] ISO/IEC 2022:1994, Information technology – Character code st ructure and extension techniques

[2] European Computer Manufacturers Association: International regist er of coded character sets to be used with escape sequences.

[3] 安岡孝一、安岡素子：コンピュータ異体字典へのアプローチ, 京都大学大型計算機センター第 52 回研究セミナー報告 (1996 年 3 月), pp.43-54.

[4] 安岡孝一：CJK Tables and Character Set Tables, http://www.kudpc.kyoto-u.ac.jp/~ yasuoka/CJK.html

[5] 台湾国家標準 中文標準交換碼、CNS 11643-1992

[6] 錦見美貴子, 高橋直人, 他：マルチリンガル環境の実現、プレンティスホール出版、1996.

[7] M. Ohta and K. Handa, RFC 1554: ISO-2022-JP-2: Multilingual Extension of ISO-2022-JP

[8] HF. Zhu, et. al, RFC 1922: Chinese Character Encoding for Internet Messages.

[9] 佐藤豊: インターネット防火壁の基礎技術と応用 – DeleGate の仕組み –, コンピュータソフトウェア, Vol.14, No.1 (1997), pp.55-63.