国際符号化文字集合を補完する大規模外字フォントの制作

上地 宏一
大東文化大学 外国語学部

近年、日本の行政用文字集合の整備が大きく進んだ。この文字集合は人文科学の分野でも有用と考えられる。その一部はパソコンでも利用できるようになったが、全体が利用できるようになるのはまだ先となる。そこで本稿では、この大規模文字集合を外字フォントとしてまとめ、当用としてパソコンで自由に利用できることとした。

Production of large-scale user-defined character font which complements the international coded character set

Koichi Kamichi
Faculty of Foreign Languages
Daito Bunka University

In recent years, development of administrative character set for the Japanese advanced greatly. This character set is also useful in the field of humanities. The part of the character set is now available on a computer, it takes a long time yet that the whole character set is to become available.

In this paper, I summarized this large character set as user-defined character font, can be used freely on a computer.

１．コンピュータ漢字処理の現状

1990 年代以降、国際符号化文字集合（ISO/IEC 10646[1]および Unicode[2]、以降 UCS と記す）の制定および拡張により 7 万字種を超える多漢字処理環境が実現した。文字コードを実際に利用するためには、OS やアプリケーションにおける文字コードへの対応およびフォントの実装（提供）が必要であり、従来ここにタイムラグが生じていたが、Windows をはじめとする OS における文字コードへの対応および、自由なフォント「花園フォント[3]」をはじめとするフォントの整備により、Ext. D 集合を含むすべての UCS 収録漢字が使えるようになった。また Unicode の Ideographic Variation Database[4]（以下 IVD と記す）による異体字グリフについても Ideographic Variation Sequence[5]（以下 IVS と記す）への対応やフォントの提供により、一般的なパソコンでも利用できる環境が整いつつある。

文字コードに収録される漢字集合については、現在も UCS への追加提案の議論が継続中である[6]ほか、自由な漢字字形データベース「グリフウィキ[7][8]」には大量の漢字・異体字データが登録され続けている。

２．日本における大規模漢字集合の整備

UCS 収録の漢字集合とは別に、日本では行政で利用される漢字を中心とした大規模漢字集合が以下のように公的機関により複数規定された。これらの漢字集合は基本的には行政処理用としての人名や地名の漢字・異体字の集合であるが、見方を変えると日本の国字なども多く含まれ、人文科学分野におけるデジタルテキストに対しても有用であると考えられる。

A) 戸籍統一文字（法務省）
B) 住民基本台帳ネットワーク統一文字（以下住基統一文字と記す）（総務省）
C) 登記固有文字（法務省）

また以上の漢字集合をまとめたものが「汎用電子情報交換環境整備プログラム[9]（以下汎用電子と記す）」（日本規格協会、国立国語研究所、情報処理学会）によって整備された。その成果を実際にコンピュータで活用できるようにするという目的で「文字情報基盤整備事業[10]」（情報処理推進機構）が実施され、各種データおよびフォントが公開された。UCS および IVD 収録の漢字集合と比較すると 7 千字程度が新たな文字として定義されることになる。このデータはインターネット上に公開され、またフォントも配布されているが、現状では新たな追加となる 7 千字についてはごく一部のアプリケーションでのみ利用可能な状態である。最終的にこの中の多くの漢字が UCS および IVD に収録されると予想されるが、実際にパソコンで使えるようになるのは数年後になると考えられる。

３．現状の問題点と提案

以上をまとめると、現状の問題点として以下の3種の漢字集合については、一般的なコンピュータで扱うことが困難である。

A) IVD（UCS および IVD 同士の重複を除く 8 千字程度）: OS、アプリケーションの対応の遅れ
B) 汎用電子の追加 7 千字 : UCS、IVD への収録待ち
C) 文字情報基盤整備事業の整備対象外の登記固有文字（9 千字程度）: 扱いの決定待ち、UCS、IVD への収録待ち

これら 3 種の漢字集合は基本的に重複しないため、単純合計で 24,000 字程度が「定義・整理されたが使えない」状態にある。

そこで、本論文では暫定的な形態として外字（ユーザー定義）領域にこれらの漢字集合を割り当て、フォントを公開することによってコンピュータで「使える」状態にすることを提案する。ここで述べる「使える」状態とは、最低限字形を表示・印刷できることであり、外字特有の問題である「検索」への非適応は考慮しないが、後述の通りデータベースとの組み合わせにより対応することが可能である。

４．2 種類の手段の検討

24,000 字を外字フォントとして整備する際、以下の 2 種類の手段が考えられる。これは UCS や IVD との重複をどのように扱うかが異なる。

A) UCS や IVD との重複、および IVD 同士の重複をすべて無視し、漢字集合すべてを外字として含める

この方法は、たとえば戸籍統一文字は 55,000 字、住基統一文字 2 万字、登記固有文字 1 万字、IVD は 28,000 字あり、それらを別々に扱う。それぞれの文字集合には ID 番号（およびそれに相当するもの）が付与されている。仮に戸籍統一文字、住基統一文字、登記固有文字の順に外字コードポイントを設定するとして、戸籍統一文字は以下の単純な式でコードポイントを求められる。

コードポイント　＝　基底コードポイント＋　戸籍統一文字番号　÷　１０

例 : 戸籍統一文字 000790  →U+F004F（基底コードポイント U+F0000 の場合）

住基統一文字は連番に相当する ID がないため、コードポイントの隙間を詰める形で収録することになる。登記固有文字については戸籍統一文字と同様の連番に相当する ID を持つため、単純な式で求められる。

この方法のデメリットとしては、同一字形に対して複数のコードポイントが存在し、それは UCS とも重複することになる（下例）。このため、実運用において字形は全く同じであるが複数のコードポイントにわたって収録される文字が大量に発生し、混乱が予想される。また、想定している文字集合の文字数の合計が 65,536 字を超えるため、現在のフォントの仕様では 2 つ以上のフォントに分割して収録せざるを得なくなり、不便となる。

例： 
UCS : U+4E00
戸籍統一文字 : 000010 番（UCS と重複）
住基統一文字 :（収録対象外）
登記固有文字 :（収録対象外）

例： 
UCS :（未定義）
戸籍統一文字 :（未定義）
住基統一文字 : J+BBE7
登記固有文字 : 01087520 番（住基と重複）

B) 重複は除いて 24,000 字の外字として収録する

この方法は既存の UCS と重複しない純粋な追加漢字のみを外字として追加するものである。このため、外字のコードポイントから計算で元の漢字集合の ID 番号を導くことは不可能であり、すべての文字についてテーブルによって変換することが必要となる。また、3 種類の大規模漢字集合および IVD の字形の重複の判定は、それぞれの漢字集合における包摂基準が異なるため、非常に困難である。

当初フォント製作の単純化にもつながる A 方式を検討していたが、重複する字形が異なるコードポイントに複数存在することが将来的に混乱を招きかねない懸念を考慮し、B 方式を採用することとした。

５．外字フォントの制作

24,000 字のうち、登記固有文字の一部（5 千字）を除いてすべてグリフウィキに字形が登録されている（図１、２，３）。このデータを用いてフォントを制作することとした。UCS には連続する外字領域が 3 区画用意されている。このうち第 0 面のコードポイント U+E000 から始まる領域は従来の外字フォントでも使われていることが

多く、また最大で 6,400 字しか収録できないため適当ではない。U+F0000（第 15 面）および U+100000（第 16 面）から始まる領域のいずれかが候補となるが、これらの外字領域を利用する有名な外字フォントはないため、番号が小さい方の U+F0000 から始まる領域を利用することにする。

字形の重複の判定は、単純にグリフウィキのデータベース上でそれぞれのグリフが同定されているかどうかで行う。片方が他方のエイリアス（参照・リンク）グリフとなっている場合は同定されているとみなす。この同定は後述の「文字情報基盤整備事業」で公開されているデータ上における同定とは粒度が異なる。一般的にグリフウィキの方がより細かい差異を別字形とみなしている。グリフウィキではユーザーが差異ありとみなした場合は分離することが可能であり、厳密な包摂ルールは存在しえないため、文字によっては差異とみなし、または同定するといったことが起こる。データは常に更新可能なため、字形が固定されないという問題点もある。このため、実際の使用にはフォントとその版の情報が重要となる。つまり、今後この外字フォントが更新されていく過程で、古い版では同定されていた２つの別の集合に含まれる漢字同士が、新しい版では区別され、別の外字コードポイントが与えられるといった可能性や、またその逆も考えられる。これは実際の利用において問題になりうるため、引き続き解決法を検討していく必要がある。

例：グリフウィキでは差異があるとみなし、「文字情報基盤整備事業」では同定するペア（左部品の 2 本の縦棒が上部の左払いと融合しているか（上）、独立して接続しているか（下）の違い）

戸籍統一文字：183440 番 

IVS：U+6B49,U+E0101 

対象となる各文字集合を、IVS、戸籍統一文字、住基統一文字、登記固有文字の優先順位において同定を行い、重複を取り除いた結果、表１のような文字集合からなる外字フォントとなった。なお、IVS は IVS 内での重複についても取り除いており、元々27,724 字あるものが 7,882 字となっている。

表１　外字フォントの各ソースと文字数

集合ソース	文字数
IVS	7,882
戸籍統一文字	10,375
住基統一文字	1,443
登記固有文字	943
合計	20,643

６．属性データ・ツールの整備

先述の通り、複数の大規模漢字集合を字形が重複しない集合にまとめて外字フォントに割り当てるため、漢字集合と外字フォントの対応関係をデータとして整備する必要がある。このデータはインターネット上でデータベースとして公開することとする。具体的には、大規模漢字集合の ID をキーワードとして検索すると、外字フォントに収録されているかどうかを提示し、収録されている場合はコードポイントを、収録されていない場合は、どの文字と同定されているかの情報を提示する。「文字情報基盤整備事業」で公開されているデータを利用して、漢字の読みや画数などからも検索できることも望ましい。外字の検索時には当該漢字をデータベースで検索し、コードポイントをコピーアンドペーストで転記するなどして実行することになる。

また、この外字フォントは字形の表示・印刷ができればよいというものではなく、将来的な情報交換の保証にも対応できなければ不十分である。外字フォントに含まれる漢字が将来 UCS や IVD に収録された場合には、外字コードポイントからあるべきコードポイントに変換（移行）できることが重要である。IVD を除き、現状では外字集合のそれぞれの漢字に対するコードポイントは未定のため変換ツールを作成することはできないが、たとえば以下のようなメタ記述で外字コードポイントと相互変換できる変換ツールを用意し、将来的にデータが無駄にならないように努力することを意識すべきである。

メタ記述の例（グリフウィキのグリフ名に同じ）：

<koseki-000010>：戸籍統一文字 000010 番
<juki-ad01>：　　住基統一文字 J+AD01
<toki-01000500>：登記固有文字 01000500 番

このように、単に外字フォントを公開するだけでなく、関連するデータやツールを合わせて公開することにより、フォントを有用・有効なものとすることができる。

７．今後の展望

現在、収録対象となる登記固有文字の一部を除く 20,000 文字のデータ最終確認を行っている段階であり、近日中にフォントの公開を行う。またデータベース検索システムとツールの開発にも着手している。本論文で述べた外字フォントを、すでに公開している「花園フォント」と組み合わせて使うことにより、10 万字種の漢字処理環境が実現できる。

現在は行政用漢字（異体字）を主体とした外字集合となっているが、将来的には日本以外も含めた漢字文化圏の UCS 未収録漢字集合についても収録対象を広げたいと考えている。特に、現在 UCS の収録が検討されている漢字を先行して収録する方向性を検討中である。この場合 UCS に収録された段階で外字領域のコードポイントを適切な UCS コードポイントに変換するツールと合わせての開発となる。UCS への追加収録の議論から収録、そして OS やアプリケーションの実装までに数年かかるタイムラグを補う役割を担えるものと思われる。

最終的には 2 つの外字領域の合計 13 万字まで収録が可能であり、いわゆる「漢字青天井問題」もおおかた解決するのではないかと期待している。

また TEI のように文字コード外字をメタ記述で処理するような別体系への適応も検討課題である。

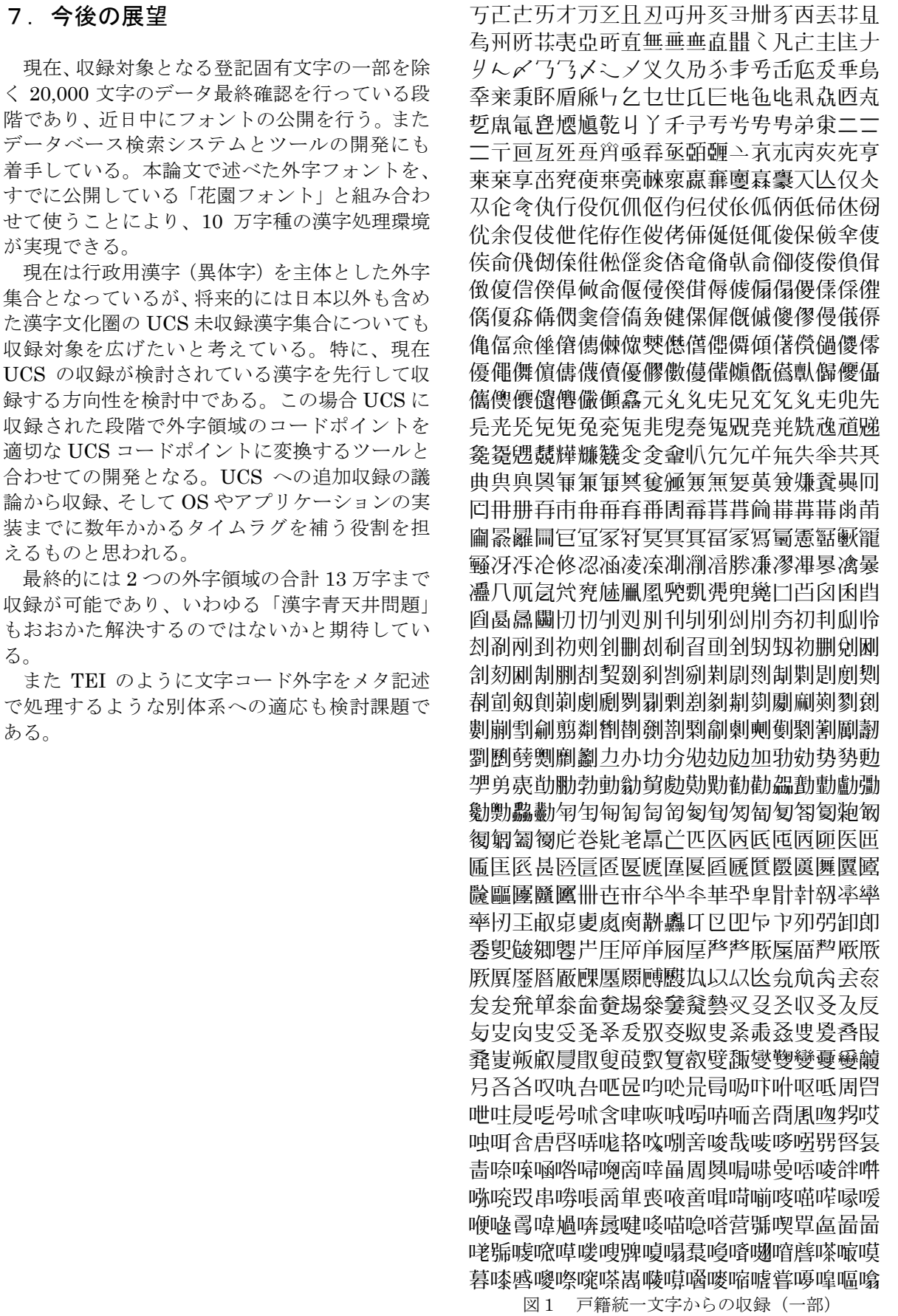

図１　戸籍統一文字からの収録（一部）

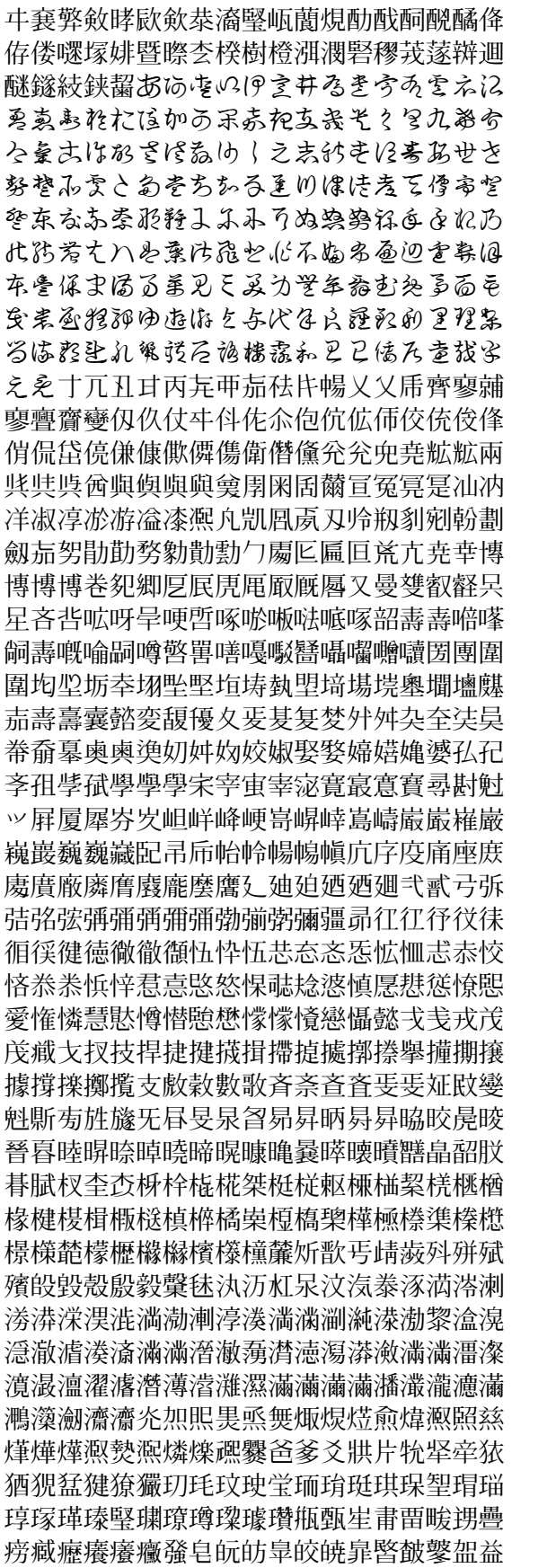

図２　住基統一文字からの収録（一部）

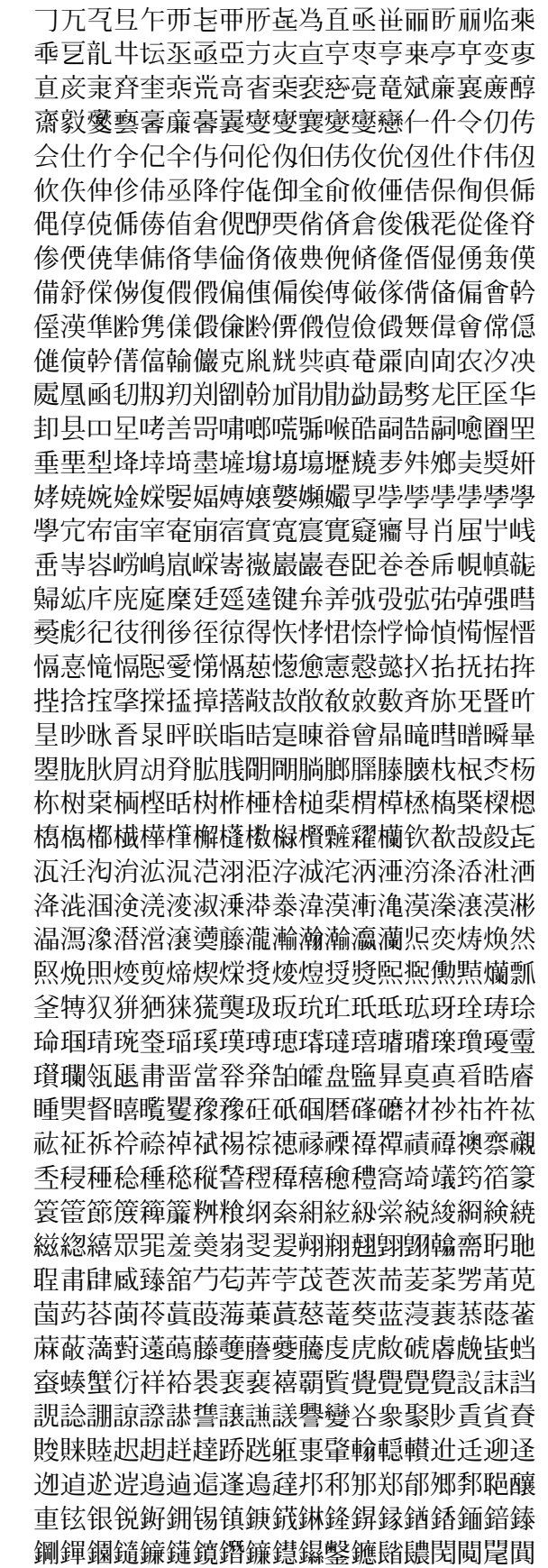

図３　登記固有文字からの収録（次ページへ続く）

陈陈陶暉隠隡粿雋睢韋糧齇飆朗餡餡餾餜饑
䪼馬馰駒駚騤驔騺驥髭鯉鮰鳥鴉鵐鶬鶳鸍齡
麠麭麯昂鼎鼎鼻齊齊齋霻鹹鹺龍龜鼇鼇鼈ク
巨梾梾桜渋しひ(平)卡忠B^{2}金
図３　登記固有文字からの収録（前ページからの続き）

謝辞

本研究は JSPS 科研費 23320010, 23320077, 22700262 の助成を受けたものです。

参考文献

1)　International Organization for Standardization：ISO/IEC 10646:2011 Universal Coded Character Set (UCS), (2011).

2)　The Unicode Consortium：Unicode 6.2.0, 入手先〈http://www.unicode.org/versions/Unicode6.2.0/〉（参照 2013-09-01）.

3)　GlyphWiki Project：花園フォント，入手先〈http://fonts.jp/hanazono/〉（参照 2013-09-01）.

4)　The Unicode Consortium：Ideographic Variation Database, 入手先〈http://www.unicode.org/ivd/〉（参照 2013-09-01）.

5)　The Unicode Consortium：Unicode Technical Standard #37, （IVS の技術規定）, 入手先〈http://www.unicode.org/reports/tr37/〉（参照 2013-09-01）.

6)　川幡太一：UCS における漢字拡張 F の標準化動向, 漢字文献情報処理研究, No.14, pp.4-7（2013）.

7)　GlyphWiki Project：グリフウィキ, 入手先〈http://glyphwiki.org/〉（参照 2013-09-01）.

8)　上地宏一：漢字グリフ管理 Wiki システム（GlyphWiki）の構築, 人文科学とコンピュータシンポジウム論文集（じんもんこん 2007）, pp.237-244（2007）.

9)　日本規格協会、国立国語研究所、情報処理学会 : 汎用電子情報交換環境整備プログラム成果報告書, (2009).

10)　情報処理推進機構：文字情報基盤整備事業について, 入手先〈http://mojikiban.ipa.go.jp/index.html〉（参照 2013-09-01）.