Vol.2015-IFAT-118 No.5
Vol.2015-DD-97 No.5
2015/3/30

画像分解による『殷墟卜辭綜類』掲出字頻度分析

鈴木俊哉[†1]　鈴木　敦[†2]　菅谷　克行[†2]

甲骨文字のデジタル化において、文字符号化して問題ないか、あるいは画像として扱うべきかは、拓本資料の鮮明さや掲出例数を考慮して判断しなければならない。我々は印刷物として公表されたデータベースである『殷墟卜辭綜類』と『殷墟甲骨刻辭類纂』に基づいた検討をすすめているが、全て手書き資料であり、また、そこに模写された文字の集合も明確ではないため、文字認識的な手法をとることができない。本研究では、掲出例数を概算するため、両書に共通するレイアウト構造をもとに模写テキストを画像分解する方法を検討した。両書の画像分解精度が大きく異なる結果が得られたが、この原因は両書の編集・出版方針の違いによると推測される。また、本研究の手法の適用範囲についても報告する。

Glyph Appearance Frequency Estimation of "Inkyo Bokuji Sourui" by Image Decomposition

suzuki toshiya[†1]　Atsushi Suzuki[†2]　Katsuyuki Sugaya[†2]

In the digitization of the Oracle Bone materials, the criteria to digitize as “coded text” or as an image should be decided with the consideration about the legibility of the source materials and the “glyphs” on the materials are sufficiently popular to interchange with the stable identity. For the character encoding in ISO/IEC 10646, once Japanese experts proposed to select the representative glyphs by the frequency of the contexts listed in the corpuses, like “殷墟卜辭類纂” or “殷墟甲骨刻辭類纂”. In this report, we estimated the frequency by automatic image decomposition method. The representative glyphs with the frequencies with the examples more than 10 are estimated about 850. This is further smaller than the estimation by the cross section of 2 corpuses.

1. はじめに

1.1 古漢字の国際標準文字符号化というアイデア

ISO/IEC JTC1/SC2/WG2 に対して、中国が ISO/IEC 10646(いわゆる Unicode 規格の基礎となる文字集合規格)[1]への古漢字が提案されたのは 2003 年のことである[2]。この提案が承認され、台湾の専門家も参加して甲骨文字の選字作業が開始され、約 10 年が経過したが、この選字作業は候補としてのグリフ集合を選定しきれず、作業部会も ISO 傘下の組織としては解散した[3][4]。

日本の専門家は、この選字作業の問題点として、文字符号化によって達成すべき具体的なユースケースが不明であり、甲骨文字の先行研究における文字整理手法や文字の同定基準(何をもって同一の文字とするか)が判然としないまま作業を開始していること、さらに、先行研究に見える整理結果ではなく、甲骨拓本資料からの文字採集をするとしたにもかかわらず、網羅的に採集するプロセスになっていないため、文字集合としての一貫性がとれないことを指摘したが[5]-[14]、この指摘に対応したプロセス設計はなされず、結局作業部会が解散されることとなったものである。

1.2 甲骨文字のデジタル化に見える近年の傾向

開発手法が改善されなかった原因のひとつに、中国や台湾の専門家の発想として「特異な文字が漏れなく収集されていればいるほど、良い文字集合である」といった傾向があった。この傾向のため「先行研究には採集漏れがあるため新たに文字採集を行う」という判断を見直すことができなかったのである。この傾向は標準化に関わった専門家に特有のものではなく、1990 年以降にコンピュータを活用して香港中文大学や華東師範大学のグループが出版している甲骨文字字形集の多く[18]-[25]が「先行研究では同一視していた文字だが実際には微細な字形差が発見された」ことを強調するために見出し字を増やすという方針をとっていることから、中国や台湾の甲骨文字研究者に共通した傾向と言える。

しかし、「同一視されていた文字に実際には微細な字形差が発見された」からと言って、その文字が特異な意味があり、一方を検索している場合にもう一方が混入することを避けなければならないのかは、出現文脈について十分な数の用例を踏まえなければ判断ができない。言い換えれば、どのような出現文脈を区別しなければならないか、という要求条件を定めなければ、それに応じた弁別基準は定まらない。現代漢字の例で言えば、「呂」と「吕」を区別すべきなのか、「兌」と「兑」を区別すべきなのか、といった基準は要求条件によって容易に変わるのである(例えば、ISO/IEC 10646 ではこれらの文字を区別しているが、既存の文字符号規格との往復互換性のためであって、既存の規格になければ統合されていた)。

残念ながら、近年の甲骨文字字形表は出現文脈を欠いたものや、あるいは出現文脈を示していても統計化されてい

†1 広島大学　総合科学研究科
Hiroshima University, Faculty of Integrated Arts and Science.
†2 茨城大学　人文学科
Ibaraki University, College of Hummanities.

Vol.2015-IFAT-118 No.5
Vol.2015-DD-97 No.5
2015/3/30

ないものが大半である。これらの情報を提供しているのは甲骨文字研究へのコンピュータ導入以前の紙媒体でのデータベースであるところの『殷墟卜辭綜類』[26](以下、『綜類』と呼ぶ)、『殷墟甲骨刻辭類纂』[27](以下、『類纂』と呼ぶ)程度である。これらのデータベースは、甲骨資料の模写を作った後、字形から出現文脈群の模写をほぼ全て辿れるような構造になっている(ただし、あまりにも用例が多いものは一部しか示していない)。この編纂方針については『綜類』の著者である島邦男が、甲骨資料の解読には帰納法的な手法を用いないと検証ができないためと述べているが、これらの考え方を引き継いだデータベースは『類纂』以降は出版されておらず、個別具体的な字形差を強調する傾向に戻っていると言える。

1.3 日本提案の考え方と本研究の問題設定

甲骨文字の標準文字符号化に際して日本から提出された代案は[16]、甲骨文字資料から直接に選字するのではなく、統計的に抽象化された先行研究として『類纂』を選び、この掲出字から、掲出例が 10 例以上あるものを同定可能として選定し、具体的な字形としてはいわゆる王朝卜辭を族卜辭より優先、さらに五期区分のうち先行する王朝の字形を優先、というプロセスで例示字形を選ぶというものであった。

『綜類』の出版からは 30 年以上、『類纂』の出版から 20 年以上経過しているが、甲骨文字の出土資料の大半は『類纂』以前に発掘されたものであり、新出資料[28)-30)]があるとはいえ、特殊な重み付けを行わない限り統計的な分析には大きな影響はないと考えられる。

しかし、『綜類』『類纂』とも紙媒体でのデータベースであり、用例数の具体的な数値は示されていないため、同定可能な文字を選定するためには用例数を求めることが必要である。しかし、『綜類』『類纂』の模写ミスなどの指摘があることを考慮すると、単純に両書の行数を数えたとしても選字作業の初期段階の一過性のデータにしかならない。標準提案のプラットフォームとしての活用を考え、本稿では以下のように問題を設定した。

- 各「用例」の画像切り出しが可能で、かつ、その文書内の出現位置が追跡可能であること。
- 各「用例」の画像切り出し情報を画像自体と分離し、公開可能であること。

本研究に先立ち、『綜類』『類纂』の中で最も見出し字形が多い部首である人部・女部の突き合せにより、以下のことを明らかにした[17]。

- 人部・女部において見出し字の 6 割程度は『綜類』『類纂』に共通している。
- 『綜類』『類纂』の突合せができない文字の大半は掲出例数が少ない文字で、10 例以上の掲出例がある文字は全て対応づけが可能であった。
- 掲出例が少ないものは後出の『類纂』での新出資料とは限らず、資料が不鮮明なものを『綜類』『類纂』で異なる形に模写しているもののほうが多い。

本稿ではこの成果を踏まえ、『綜類』『類纂』の見出し字対応づけでなく、用例数基準によって絞り込んだ場合にどの程度が安定して符号化可能かを見積もりたい。もし『綜類』『類纂』で大幅な違いがあれば、両書で模写が大幅に違っている可能性があり、さらなる精査が必要となるからである。

2. 本研究での手法

2.1 先行研究

文書画像からの行抽出問題は、OCR の前処理としての領域分割問題として研究が進められてきた。手書き文書の領域分割方法としては足立らの研究があるが[31]、近年の領域分割問題に関する研究は活字またはデジタルフォントによる印刷物の精度を上げるか、あるいは景観中の文字のような背景のノイズが非常に強い状況での認識精度を上げるといった問題設定に移りつつあり、手書き文書の領域分割手法の実装展開は進んでいない。たとえば、文書画像に対する検索・注記プラットフォームである SMART-GS も文書画像に対する行分割に関しては十分に自動化されておらず、手作業での処理が安定解であるとしている[32)33)]。しかし、『綜類』は 500 ページ 92000 行、『類纂』は 1300 ページ 105900 行を超える文書であり、たとえば人海戦術的な方法で実施したとすると品質の確認が困難である。そこで本研究では画像処理的な手法での処理を評価した。

2.2 本研究の手法

本稿で対象とする『綜類』『類纂』の本文レイアウトは図 1 のようになっている。

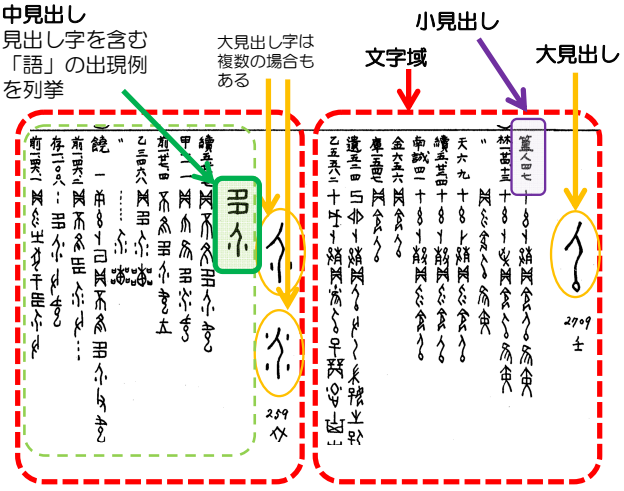


図 1:『綜類』のカラム内部構造

論理的に区別しなければならないものとして、以下のものがある。

大見出し: 例字字形(複数)

中見出し: 例字字形の前後一文字を付加し、「語」と思われるパターン

Vol.2015-IFAT-118 No.5
Vol.2015-DD-97 No.5
2015/3/30

小見出し: 出現文脈を採集した資料名
本文: 出現文脈(行折り返しあり)

この中で、大見出しと小見出しおよび本文の文字サイズは大きく異なるが、大見出しと中見出しの文字サイズには目立った差がなく、「大見出しの後の中見出し」という大域的な前提がなければ区別が難しい状況も少なくない。大見出し字数は『綜類』では 3868 字、『類纂』では 4499 字あり、自動化が望ましい数ではあるが、中見出しは大見出し 1 項に対して 10 項近く掲出されることもあるため、大見出し字と中見出し字が混在した状態で切り出すと、そこからの大見出し字抽出は非常に困難である。これを踏まえて大見出し字は手作業により分割する。

見出し字以外の部分については、以下の手順によって画像分割を行った。

① ページ単位の画像をカラム別に分割する。『綜類』では 1 ページあたり 5 カラム、『類纂』では 1 ページあたり 4 カラム(うち 2 カラムは隷定テキストである)が含まれるが、ページ単位で始点・終点を定めた線分として罫線を検出した場合、スキャン時の歪みのためにごく一部しか検出できない場合が多いため、カラムごとに実施する。

② カラム別画像に対して罫線を基準に斜行補正を行う。罫線検出には OpenCV[34]を利用し、角度のみを検出した。ページ画像に対して回転処理を行うと鮮明さが損なわれるため、元画像を SVG 形式に変換し、CSS transform を与えることにより表示系に回転させるものとした。以降の処理は、ラスタ画像に関わる部分は全て元画像の座標系で実施している。検出例を**図 2** に示す。カラム分割しても罫線を完全には抽出できず、断片的な罫線から外挿してカラムの四隅を求め、それによってカラムを切り出した。

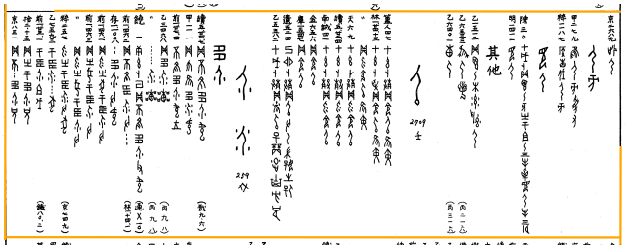


図 2: カラムに対する罫線検出の例

③ カラム別に画像をアウトライン化し、各パスの外接矩形をもとに閾値より小さなパス(汚れなど)、閾値より大きなパス(罫線断片など)を除外する。アウトライン化には potrace[35]を利用した。**図 3** に文字構成図形の検出例を示す。マゼンタで示しているのは除外されたものである。

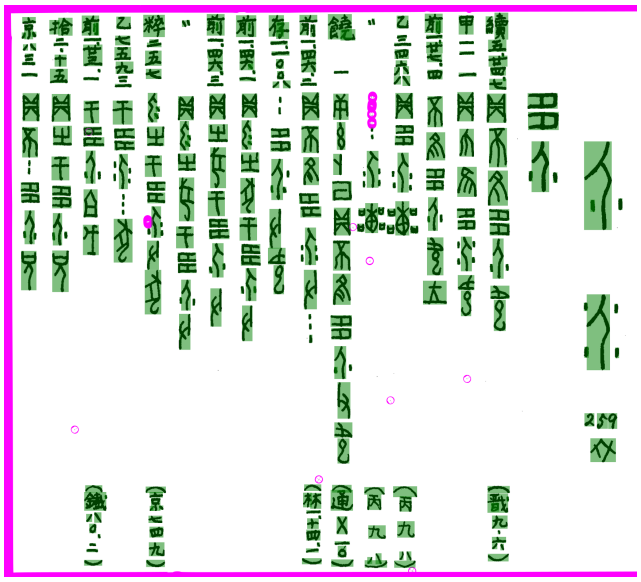


図 3: 文字構成図形の検出の例

④ 手順③で除外されなかった図形(文字あるいは文字を構成する部品図形)の外接矩形について、ある矩形の中央が上下に隣接する矩形の幅に収まる場合、同じ行に属するパスとみなす。ある行に属するパスを全て含む外接矩形を行の基本矩形とみなす。行の検出例を図 4 に示す。手順④ではどの行にも属しなかったパスについて、これと重なりのある行のうち、もっとも基本矩形との重なり面積が大きい行に属するとする。手順⑤で属した矩形に関しては行基本矩形には含めない。

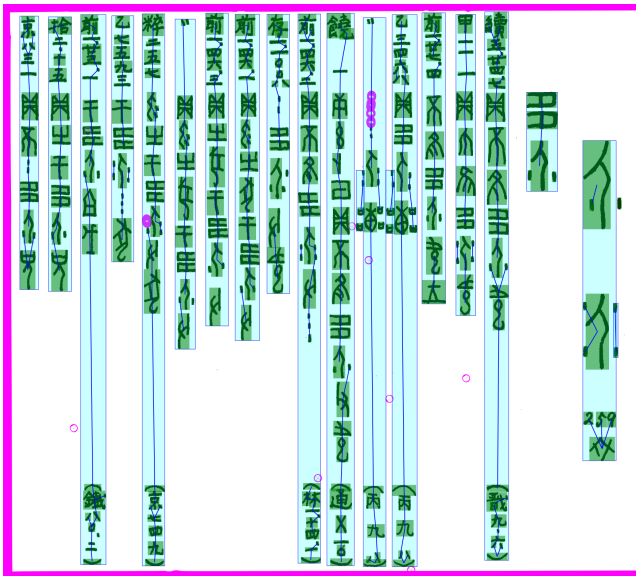


図 4: 文字構成図形による行領域の検出例

⑤ 行矩形の幅が指定した閾値の範囲に収まっているもの(600dpi でスキャンした画像に対し、80～150 ピクセルをこの範囲とした)を正常に切り出された行とみなす。また、行開始位置が指定した閾値(『綜類』ではカラム上端より 150 ピクセル以内、『類纂』では 250 ピクセル以内)に収まっているものは資料名を含

Vol.2015-IFAT-118 No.5
Vol.2015-DD-97 No.5
2015/3/30

む行であると判断する。正常に切り出されなかった行は、たとえばある行に離れた点がある文字が一つだけ含まれているが、上下にそれを含む幅の文字がないために「点だけを含む行」として分割されてしまうような事例である。本稿では純粋に行を構成する文字だけを切り出すことを目標とはしておらず、行数を求めるため、致命的な問題とはならないが、行分割をもとにSMART-GSなどで画像検索を目指そうとする場合にはこれらの断片の再構成も課題となる。

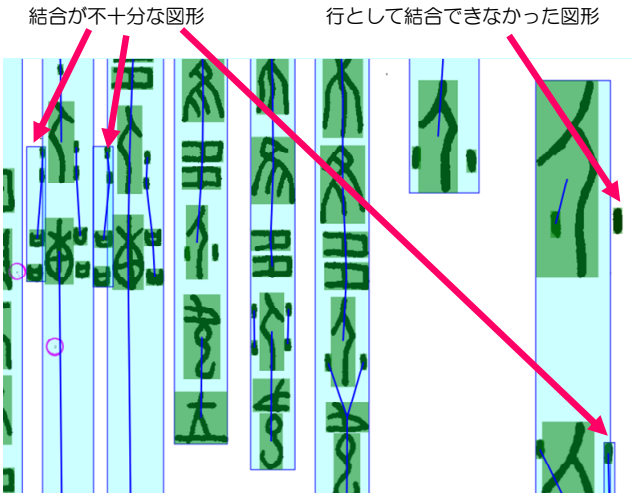


図 5: 行構成が不完全となる例

以上により分割した行について、以下のように例字字形ごとのグループに分ける。

① 各用例群の前に示される例字字形の領域情報を指定する。

② ある例字字形の行の後、「同じ行に属さない例字字形」が出現するまでのカラム内の行は全て例字字形の掲出例と解釈する。

ここで、例字字形の切り出しを手作業で行った理由を述べる。

3. 結果

3.1 検出結果

見出し字グループを掲出例数に基づいて 10 例以上、10 例未満 2 例以上、1 例のみに分類した結果を**表 1** に示す。

	『綜類』	『類纂』
10 例以上	846	876
2〜9 例	1324	1194
1 例	1157	1526
総グループ数	3327	3596

表 1: 画像分解に基づく掲出例数

見出し字グループ総数の差が 269 あるのに対し、10 例以上の見出しグループの差は 30 であり、当初方針の「10 例以上掲出例があるものを安定した文字同定が可能とみなす」という判断は有効であると考えられる(総グループ数に対する比率で換算すると、『綜類』では 25%、『類纂』では 24%となる)。各グループを例数の順に並べ、その例数および全体に対する累積の割合(当該グループより多くの例数を含むグループを集め、その例数の合計の全体例数に対する比率を示したもの)を**図 6** に示す。

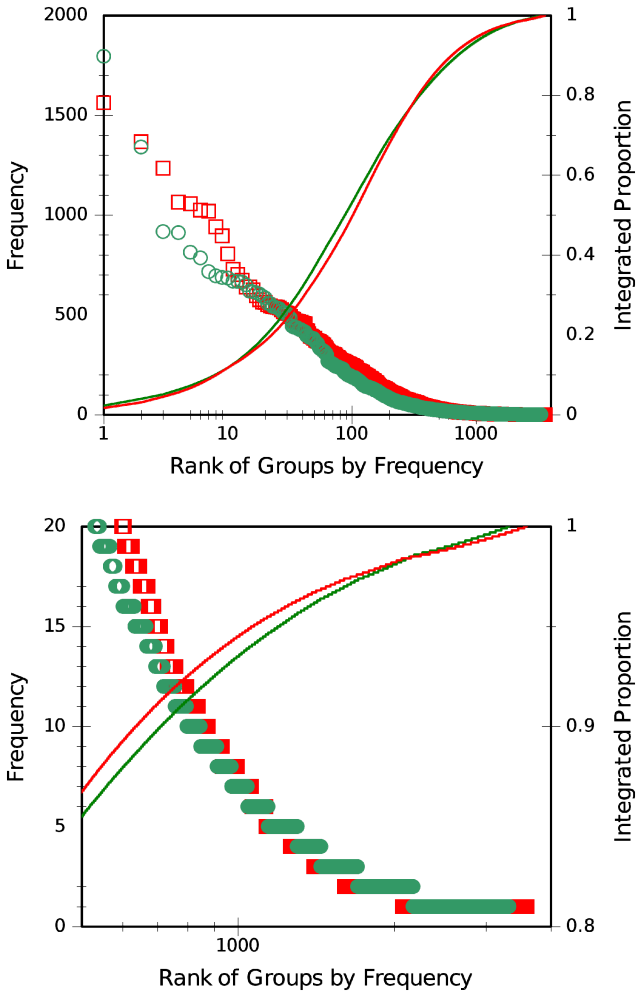


図 6: 例数順の掲出例数および累積割合

緑○: 『綜類』掲出例数、緑線: 『綜類』累積割合

赤□: 『類纂』掲出例数、赤線: 『類纂』累積割合

3.2 誤分解について

本研究の手法で行分割を行った結果、誤分解の可能性がある行について**表 2** に整理した。行幅(w)が単純平均(w_{avg})に対して 2 倍以上の幅を持つ行($w > 2w_{avg}$ となるもの、隣接する 2 行を結合したおそれがあるもの)は、『綜類』では 778 グループ、『類纂』では 114 グループであった。また、平均的な行幅の半分以下のもの($w < w_{avg}/2$ となるもの、本来 1 行に収まるべきものが 2 行に分割されてしまったおそれがあるもの)は、『綜類』では 1160 グループ、『類纂』では 960 グループであった。厳密には、「画像ではたしかに分割できる筈だが、手法の限界による誤分割」「画像で既に接触しているため、何らかの前提知識がなければ分割・結合できない誤分割」などを区別しなければならないが、簡単のため、表では一括して示した。行数ベースでは過剰分解が疑われ

るものは『綜類』で 5%、『類纂』では 3%、過剰結合が疑われるものは『綜類』で 1%、『類纂』では 0.1%であり、掲出例数が 9 例か 10 例という境界領域では問題になるが、**図 6** に示した傾向から判断して、この誤差によって増減する「10 例以上の見出し字」の数は 100 個程度であろうと考えられる。

	『綜類』		『類纂』	
	行数	グループ数	行数	グループ数
$w < w_{avg}/2$	4992	1160	3485	960
$w > 2\ w_{avg}$	978	778	124	114
総数	92988	3327	105955	3596

表 2: 誤分解のおそれがある切り出し行の数

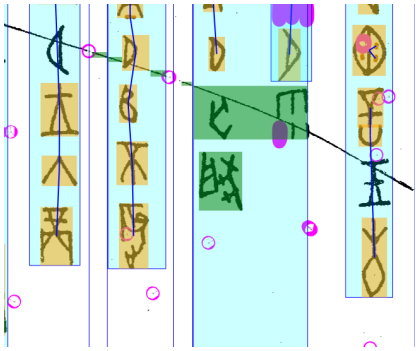
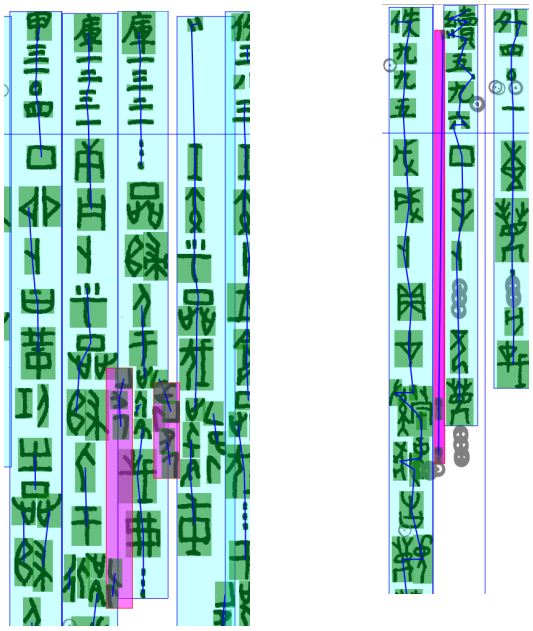


図 7:『類纂』の印刷汚れによる模写字形の誤認識例



模写が入り組んでいる例　　小見出しと模写が入り組んでいる例

図 8:『綜類』の行入り組みによる誤認識例

『綜類』『類纂』の結果を見た場合、特に目立つ差は過剰結合が疑われるものが『綜類』で非常に多いという点である。これは、『類纂』は模写のみのデータベースをまず作り[26]、これの切り貼りによって作業したため行間が比較的開いており、過剰結合の原因は印刷に際して版面に散らばっている汚れに起因する(**図 7**)のに対し、『綜類』は同様の編集作業を行った後に著者による清書が行われ、行間が非常に狭く、互い違いに入り組んでしまっている(**図 8**)状況が多いことに原因があると考えられる。

4. まとめ、今後の課題

本稿では、『綜類』『類纂』の見出し字に対して、本文で掲出される例の数を画像処理的手法で検出することにより、文字の同定基準が安定していることを期待できるものを選定した。その結果、両書とも 850 字程度が 10 例以上の掲出例数を持っていることがわかった。今後、まずこれらの例字字形を付き合わせることにより、例字字形の選定を進めることが標準化提案に向けて必要な課題である。

また、出現文脈が 10 例未満のものは、先行研究では 3 割程度と見積もられていたが、これよりも遥かに多く、割合では逆転しているおそれが高い。これらの文字を標準文字符号に含めるべきかどうかの検討も必要であろう。特に、非常に頻繁に使われ、かつ、機能的な議論が不要なもの文字(干支、数字)に関しては『綜類』『類纂』は節録としており、全ての掲出例を示していない。これらの文字が 2～9 例のグループに含まれている可能性があり、より詳細な分析が必要である。

最後に、『綜類』『類纂』は見出し字の作り方の指針が異なり、『綜類』では共通化･抽象化をはかっているのに対し、『類纂』では具体的な字形に近づけているという指摘がある[5)7)]。本稿では両書の掲出例数に関して極端な差は見られなかったが、**図 6** での累積割合を詳細に見ると、掲出例が 5 例までの領域では『類纂』が急激に掲出例数を減少させているが 5 例未満の領域ではむしろ『類纂』がロングテール的な傾向を示しており、見出し字設計の方針の違いが影響している可能性がある。この領域の分析も異体字の取り扱い指針の検討のためには必要であろう。

謝辞　本研究は、科研費基盤研究(C)課題番号 24500116、26330377 の成果を含む。

参考文献

1) ISO/IEC 10646: Information technology -- Universal Coded Character Set (UCS), International Standard Organization, Switzerland (2014).
2) 李國英, Tom Bishop: “Draft Agreement on Old Hanzi Encoding”, ISO/IEC JTC1/SC2/WG2/IRG N1014 (2003).
3) ISO/IEC JTC1/SC2/IRG: “Resolutions of IRG Meeting #38”, ISO/IEC JTC1/SC2/IRG N1870 (2012).
4) ISO/IEC JTC1/SC2/IRG: “Report from the Old Hanzi Expert Group”, ISO/IEC JTC1/SC2/IRG N1215 (2006).
5) 鈴木敦: “Input to Old Hanzi Expert Group”, ISO/IEC JTC1/SC2/WG2/IRG N1346 (2007)
6) (Japan NB): “Classification of oracle bones based on prior researches on their usages” ISO/IEC JTC 1/SC 2/WG 2/IRG N1424 (2008)
7) 鈴木敦: “Questions on the policy of old hanzi expert group works” ISO/IEC JTC 1/SC 2/WG 2/IRG N1522 (2008)
8) 鈴木敦: “対古漢字中的甲骨文字進行符号化処理的問題点”,

南方文物, Vol. 67, No. 163, pp.124-129 (2008)
9) 鈴木敦: “Old Hanzi における甲骨文字符号化作業の問題点と金文・列国文字符号化作業への影響”, 東洋学へのコンピュータ利用第 20 回研究セミナー (2009/03/27)
10) 鈴木敦: “論先秦文字符号化問題”, 紀念王懿榮発現甲骨文 110 周年国際学術研討会 (2009/08/14)
11) 鈴木敦: “先秦文字の符号化に関する諸要件”, 中国出土資料学会平成 21 年度大会 (2010/03/13)
12) 鈴木敦: “Concerns on Old Hanzi Activities” ISO/IEC JTC 1/SC 2/WG 2/IRG N1695 (2010)
13) 鈴木敦: “先秦文字の符号化に関する諸要件”, 茨城大学人文学部紀要人文コミュニケーション学科論集, No. 9, pp.75-84 (2010)
14) 鈴木敦: “先秦文字の符号化作業の現状と課題”, 情報技術標準 NEWSLETTER, No.89, p.2-5 (2011)
15) Adam Smith: “Comments on the work of the Old Hanzi Group towards an encoding of OBI script”, ISO/IEC JTC1/SC2/WG2 N4236 (2011).
16) (Japan NB): “Japan’s Proposal of Oracle Bone Coding Framework”, ISO/IEC JTC 1/SC 2/WG 2/IRG N1771 (2011)
17) 鈴木敦, 鈴木俊哉: “甲骨文データベースのデジタル化諸要件と作業プロセスの検討”, 東洋学へのコンピュータ利用第 24 回研究セミナー, ISSN 0910-3201, p.15-74.
18) 范麗梅, 張再興, 洪一梅, 何志華, 程少軒, 魯家亮, 李靜, 單育辰: “「數位時代的出土文獻」專輯”,中國文哲研究通訊, Vol. 21, No.2 (2011) p. 1-113.
19) CHinese ANcient Texts, http://www.chant.org/
20) 蔡世彬: “甲骨文全文資料庫--現代科技与古老文字的結合”, 中大通訊, Vol. 12 (10), No. 185, p.4 (2001-6-4)
21) Che Wah Ho: “CHANT (CHinese ANcient Texts): a comprehensive database of all ancient Chinese texts up to 600 AD”, Journal of Digital Information, Vol. 3, No.2 (2002).
22) 沈建華, 曹錦炎: “新編甲骨文字形總表”, 中文大学出版社, Hong Kong (2001) ISBN 9789629960476
23) 沈建華, 曹錦炎: “甲骨文字形表”,上海辭書出版社, Shanghai (2008) ISBN 9787532624317
24) 劉志基: “古文字考釋提要總覧”, 上海人民出版社, Shanghai (2008) ISBN 720807982X, 7208092117, 7208104700
25) 陳年福: “殷墟甲骨文摹釋全編”,線装書局, Beijing (2010) ISBN 7512002920
26) 島邦男: “殷墟卜辭綜類”, 汲古書院, Tokyo (1971)
27) 姚孝遂: “殷墟甲骨刻辭類纂”, 中華書局, Beijing (1989) ISBN 9787101004779
28) 中國社会科学院歷史研究所: “甲骨文合集補編”, 語文出版社, Beijing (1999) ISBN 9787801264961
29) 中國社会科学院歷史研究所: “殷墟花園庄東地甲骨”,雲南人民出版社, Kunming (2003) ISBN 9787222038776
30) 齊航福, 章秀霞: “殷墟花園庄東地甲骨刻辭類纂”, 線装書局, Beijing (2011) ISBN 9787512002982
31) 足立雄介, 吉川大弘, 鶴岡信治: “細線化処理を用いた手書き文章領域からの文字列分割”, 信学技報. PRMU 98(528), 121-126, 1999-01-22
32) SMART-GS マニュアル, “イメージサーチ/テキストサーチ”, http://www.shayashi.jp/HCP/SMART-GS/manual/imagesearch/imagesearch.html (2015-03-09 閲覧)
33) 寺沢健吾, 川嶋稔夫: “文書画像からの全文検索のオンラインサービス”, 人文科学とコンピュータシンポジウム, 2011 年 12 月, https://ipsj.ixsq.nii.ac.jp/ej/index.php?action=pages_view_main&active_action=repository_action_common_download&item_id=79427&item_no=1&attribute_id=1&file_no=1&page_id=13&block_id=8 (2015-03-09 閲覧)
34) OpenCV, http://opencv.org/ (2015-03-09 閲覧)
35) Peter Selinger, http://potrace.sourceforge.net/ (2015-03-09 閲覧)
36) 姚孝遂: “殷墟甲骨刻辭摹釋總集”, 中華書局, Beijing (1988) ISBN 9787101003451