様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １５ 日現在

機関番号：１５４０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１１６

研究課題名（和文）表意文字曖昧検索のための文字表現および検索手法の研究

研究課題名（英文）Study on the Identification of Ambigous Ideographic Characters

研究代表者
鈴木　俊哉（suzuki, shunya）

広島大学・情報メディア教育研究センタ・助教

研究者番号：７０３１１５４５

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：甲骨文字資料の網羅的な整理である『殷墟卜辭綜類』(島邦男, 1971)と『殷墟甲骨刻辭類纂』(姚孝遂, 1989)の見出し字を抽出して対応関係を整理した。また、画像処理的手法により各見出し字の出現頻度を調査結果と組み合わせた。『綜類』検字総表7569字と、『類纂』項目見出し字4499字から約2600字が対応可能で安定した文字域を持っており、そのうち10個以上の例数により帰納的な文字同定が可能と思われるものは900項目未満であった。また、新出資料により同定基準が変化した例は少なく、新出資料の網羅性はあまり影響がないことがわかった。同様の調査を女書に対しても行い、国際標準への提案などを行った。

研究成果の概要（英文）：We compared major 2 databases for Oracle Bone texts; "Inkyo Bokuji Sourui" (1971) and "Yixu Jiagu Keci Leizuan" (1989). We collected all indexing glyphs (7569 glyphs from "Inkyo Bokuji Sourui", 4499 glyphs from "Yixu Jiagu Keci Leizuan") and made the mapping table without the consideration of their context. The mapped glyphs that could be found in both database are estimated about 2600. However, the glyphs with sufficient popularities (found in the objects more than 10) are estimated about 900. Their identities could be regarded as stable, and ready to include the standard and stabilized core character set. Considering the comparison of 2 databases, it is foundt that the new objects found or published during 1971-1989 have no serious impact with the stability of the character identity.

研究分野：情報工学

キーワード：殷墟卜辭綜類　殷墟甲骨刻辭類纂

１．研究開始当初の背景

いわゆる Unicode のベースとなっている ISO 規格、ISO/IEC 10646 への文字の追加は、近年では広いユーザ層が活発に利用している文字の大半が収録済となった現在、追加提案が歴史的文字・少数民族文字の追加にシフトしてきているが、それらを主要な書記言語として用いるユーザが既に存在せず、文字の同定基準を解読が不完全な限られた資料で議論しなければならないという状況に対応するためのものであった。

研究開始当初に念頭においていたものは、甲骨文字に代表される古漢字(漢代の、いわゆる隷書が確立する以前の漢字を総称してこう呼ぶ。2003 年に甲骨文、金文、説文小篆の 3 つの文字の ISO/IEC 10646 への追加が提案された)、および、古彝文(1970 年代に規範化された四川省大涼山･雲南省小涼山地区の彝文字に対し、それ以外の地域での規範化されていない伝統的な彝文字を呼ぶ。これに関しても 2008 年に最初の提案が出された)であった。

２．研究の目的

本研究課題は、さまざまな表意文字において見られる、文字の同定基準が安定しないため文字集合の定義が困難な場合に、どのように文字を同定・検索するかという問題意識に基づく。本研究の目的は、第一に甲骨文字の曖昧検索のための同字・別字判定基準の検討と文字同定基準が安定している部分文字集合の選定と、第二にその手法が他の歴史的・少数民族文字への適用が可能かどうかの検討である。

３．研究の方法

研究方法としては、まず甲骨文字資料の網羅的な整理である『殷墟卜辭綜類』(島邦男, 1971)と『殷墟甲骨刻辭類纂』(姚孝遂, 1989)の見出し字を抽出して単純な図形レベルで比較し、対応表を作成する。これにより、まず「研究者に依存しない甲骨文字の例字字形」を得て、甲骨文字の研究に習熟していなくても扱える例字字形の集合がどの程度の規模となるかを検討した。

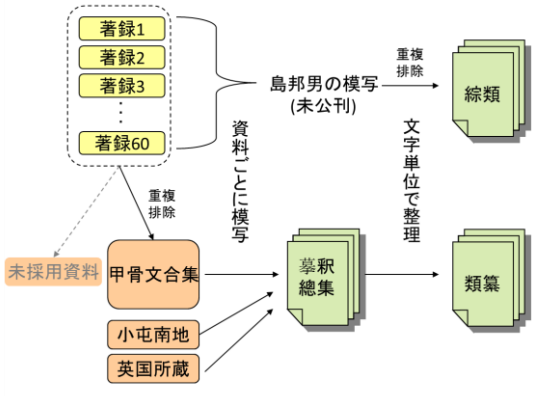

図 1:　『綜類』『類纂』の関係

『綜類』は現在の甲骨文字研究の基盤となっている拓本データベース『甲骨文合集』出版以前に様々な拓本資料を模写した上で出現文脈を字形単位で整理しなおしたデータベースである。これに対し、『類纂』は同様の整理方針であるが、『甲骨文合集』に加え『懷特氏等所藏甲骨集』『小屯南地甲骨』『東京大学東洋文化研究所所藏甲骨集』『英国所藏甲骨集』を加えたものである(図 1)、すなわち、『綜類』の文字同定基準に多数の新出資料を加えた場合にどの程度影響が生じるか、また、手法として同じであっても別の研究者が整理した場合にどの程度字形に揺れが生じるかを検討することができる。

この調査のために、文字切り出しプログラムと、対応作業プラットフォームとして JavaScript による画像リスト操作プログラムを作成した(図 2)。

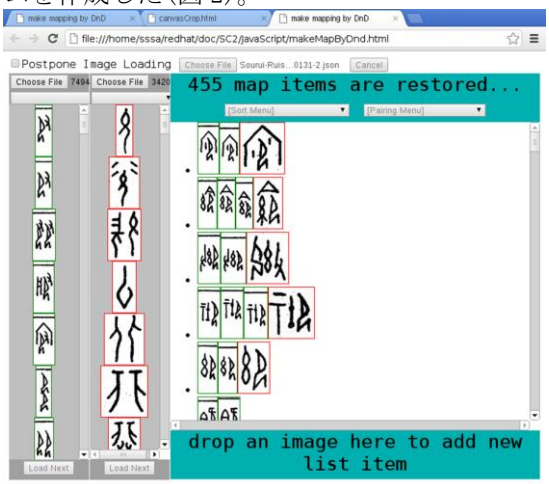

図 2:　作成した画像リスト操作プログラム

次に、各見出し字の用例として掲出される文脈数を画像処理的手法で計算し、各見出し字の出現例数を見積もった(図 3)。これにより、『綜類』『類纂』の見出し字の字形類似性だけで対応関係を見出せなかったものについて、十分な例数がありながら対応関係がとれなかったもの(図 4)を、文字の概念は安定しているが例示字形に大きな差があるものと考え、統合規則の検討材料とした。

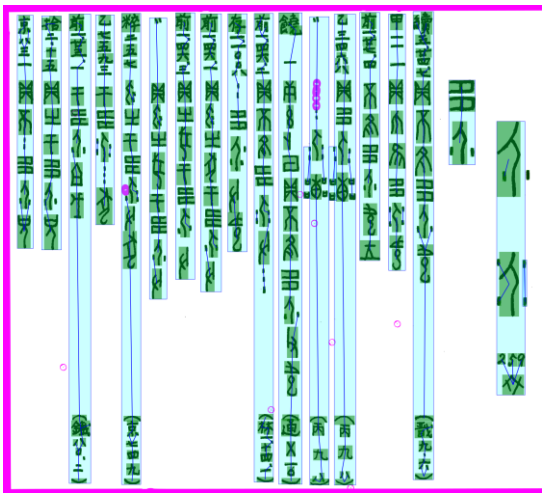

図 3:『綜類』掲出例数分析のための画像分解

また、突合せの際に同一典拠でありながら研究者による字形差が疑われるものに対し議論が可能なよう、『甲骨文合集』『甲骨文合

集補編』の掲出拓本を拓本番号ごとに画像分解したデータベースを作成した。『綜類』は『甲骨文合集』以前の原著録の名称によるので、『綜類』の著録名称と『甲骨文合集材料来源表』の著録名称の対照表を作成し、これによって『綜類』『類纂』の拓本名称を対応づけられるようにした。

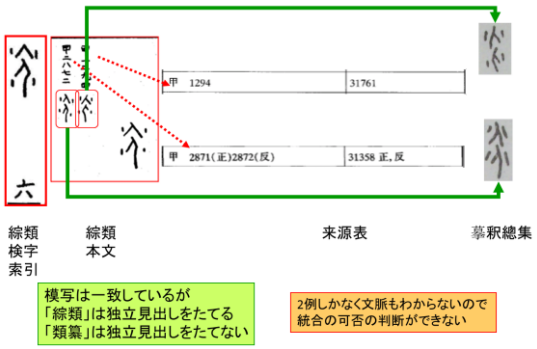

図 4:『綜類』『類纂』で統合基準にずれがある例

さらに、『綜類』『類纂』のような出現文脈データベースとしては不十分であるが、『類纂』以降に中国・台湾の研究機関において甲骨文字のデジタル化を念頭において設計された甲骨文フォントについて、その内容を調査し(図 5, 6)、文字の追加および改変の方針を調査した。これにより甲骨文字のデジタル化に関わる領域での、近年の文字同定基準の動向を推定した。

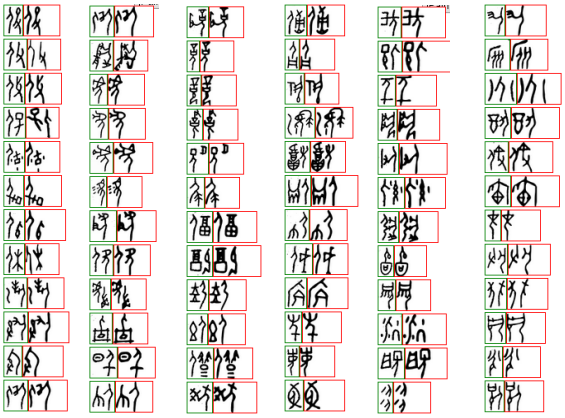

図 5:華東師範大・香港中文大甲骨フォントの比較(部分)

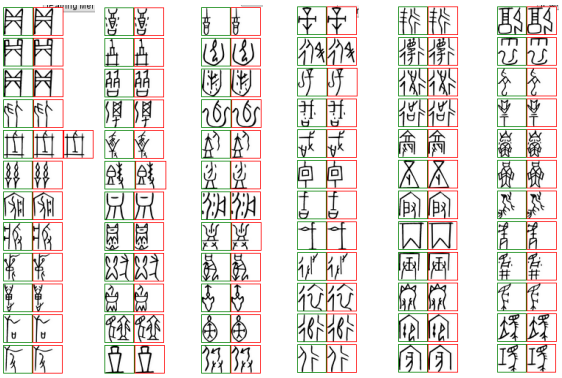

図 6: 華東師範大甲骨フォントと『類纂』の比較(部分)

4．研究成果

4．1．　甲骨文字に関する成果

初年度には、『綜類』『類纂』でもっとも収録字数が多い人部および女部について見出し字形を突合せ、約 6 割が甲骨文字の研究に習熟していなくても対応づけが可能であり、両データベース間で微細な字形差しか持たないことがわかった。また、対応づけられなかった 4 割は、出現例数が 10 例未満であり、統合可能性を議論するには不十分であることがわかった(表 1)。

	綜類 模写数	類纂は 模写せず	類纂は 異字形で 模写	類纂は 同字形で 模写	典拠不明
人部	61	3	15	35	8
女部	62	2	27	19	8

表 1: 人部・女部の『綜類』『類纂』対応関係

突合せできなかった見出し字について典拠拓本まで辿って調査した結果、その多くは同一の拓本を用いているが、不鮮明であるものや、あるいは甲骨が細かく砕けているためにどの部分を図形として 1 文字として判断するかが両書で異なっているものであった。また、一方が判読できないとして模写しなかった場合もあった(図 7, 8, 9)。

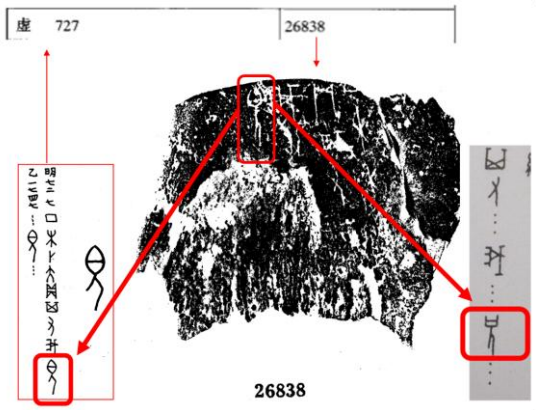

図 7:拓本が不鮮明で模写が異なっている例

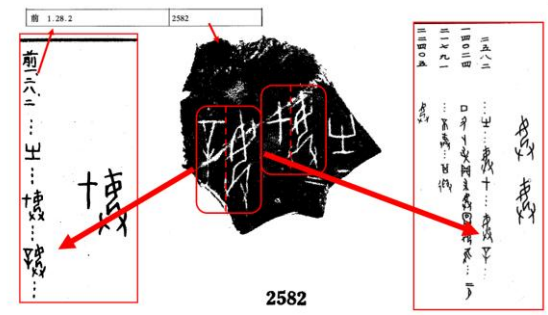

図 8: 1 文字の判定が異なっている例

また、この調査の仮定で『甲骨文合集来源表』の誤りと思われるものも見つかった(図 10)。この問題は本稿の『綜類』『類纂』対応づけの失敗にとどまる問題ではなく、資料の同定に関わるものであり、別途網羅的な調査が必要と考えられる。

図 9：一方は判読できずとした例

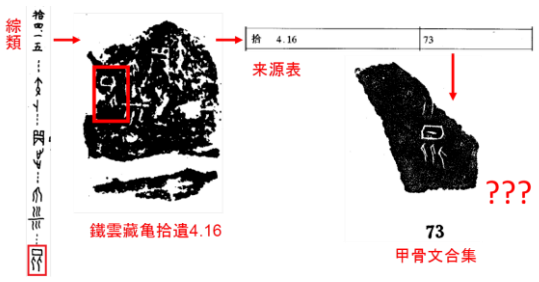

図 10:『来源表』の誤りにより、『綜類』と『類纂』の対応関係が追跡できなかった例

これらの少数出現字が文字域にどのような影響を及ぼしているかを調査するため、『綜類』と『類纂』の見出し字の対応関係調査を全項目に対して行い、また、それぞれが掲出する拓本数を分析し、両書での文字対応関係と字形の揺れの関係分析を行った。具体的には、『綜類』検字総表 7569 字と、『類纂』項目見出し字 4499 字の対応関係を調査し、対応づけができなかった約 2000 字に関して、出現例数の多寡によって検討した。例数について、『綜類』『類纂』の本文を画像処理的手法により行分解し、拓本番号部分を検出して例数の概算とした(それぞれが掲出する拓本名の数であり、厳密な意味で文字出現数や甲骨数とは言えないが、概算での文脈数と考えることはできる)。

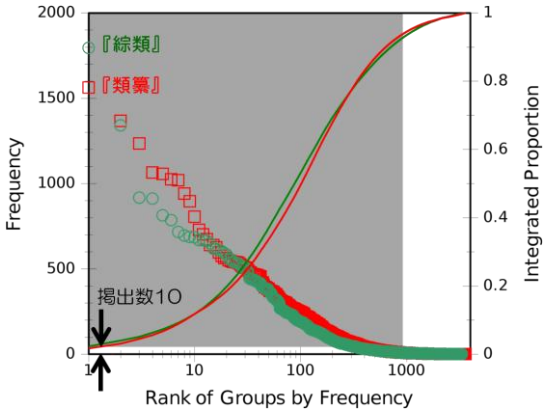

図 11:『綜類』『類纂』の項目あたり掲出文脈数

その結果、字形対応づけが困難なものの大半は 10 例未満であり、統合の可否を判断することが困難であることがわかった。具体的には、出現例数が 10 例以上の項目は 900 項目未満であり(図 11, 12)、統合の可否を用例から機械的・帰納的に判断できる甲骨文字は 1000 文字程度と見積もられることがわかった(図 13)。また、『綜類』『類纂』の比較から、従来、出現例数が少なく同定困難であった文字が、新出甲骨資料によって帰納的に同定できるようになった例は殆どなく、その意味では統合範囲を議論する材料の検討においては新出資料の網羅性はあまり影響がないことがわかった。

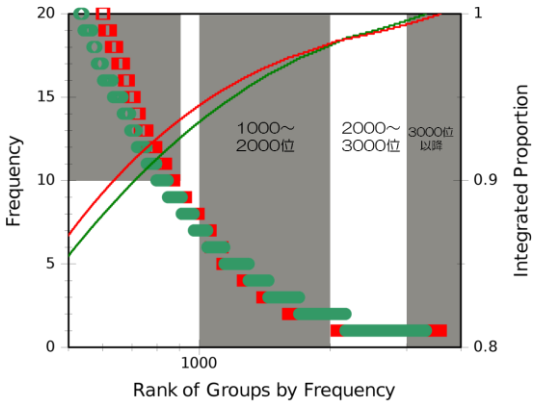

図 12：『綜類』『類纂』の掲出文脈 10 例未満の状況

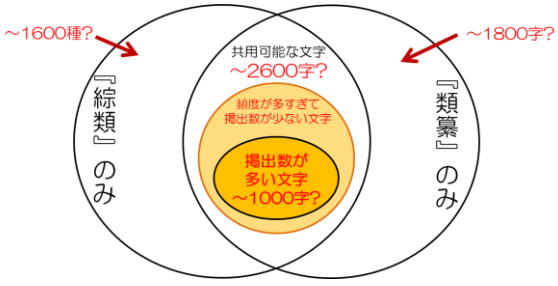

図 13：安定した文字域の項目数の推定

この画像分解の結果に関しては、著作権に影響されないよう行および見出し字のジオメトリ情報として 2015 年度中の公開を目指している。

４．２．　甲骨文以外の古漢字に関する成果

本課題の実施中、台湾と中国により説文小篆の ISO/IEC 10646 への追加が提案された。2004～2005 年に提案された際には、大徐本、小徐本、段注本、唐写本木部残巻、『説文義証』、『説文通訓訂声』、『説文句讀』から採集したグリフを整理して文字集合を決めることになっていたが、2014 年の提案では大徐本と段注本のみとなり、大徐本も藤花榭本に限定する単純化を行っていた。

この単純化の影響を調査するため、特に小徐本の字形について重点的に調査し、四部叢刊所収述古堂本、汪啓淑本、祁寯藻本の部首一覧を切り出し、續古逸叢書影印靜嘉堂本(大徐本)との比較表を作成した。その結果、祁寯藻本のみ字形が異なっている場合もあり、校訂の典拠についてより調査が必要であることがわかった。續古逸叢書影印靜嘉堂本本文、小徐本祁寯藻本本文、さらに同文書局版康熙字典欄外からの説文小篆切り出しを行い、改善提案の資料を整備した。この切り

出しに関するジオメトリ情報も 2015 年度中に公開の予定である。

４．３．非漢字に関する成果

図 14：女書集合の修正提案

本研究の手法を中国西南部江永県で使われていた女書の文字集合に適用し、ISO/IEC 10646 に提案されている文字集合に対して修正が必要であることを示した。これにより、拙速な標準化を抑止することができた。これに関してはさしかえ案の提案も行っている。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

1. 鈴木俊哉：“モンゴル語に関連した文字の標準化現況とその問題点”，日本モンゴル学会紀要，査読無，No. 45 (2015)，p. 146-147
2. 鈴木俊哉、鈴木敦、菅谷克行：“画像分解による『殷墟卜辭綜類』掲出字頻度分析”，情報処理学会研究報告，査読無，DD-97-5 (2015)，p. 1-6
3. 鈴木俊哉：“女書の標準化動向とその問題点”，情報処理学会研究報告，査読無，DD-94-5 (2014)，p. 1-7
4. 鈴木敦、鈴木俊哉：“『殷墟卜辭綜類』の部首内排列方法の分析”，情報処理学会研究報告，査読無，DD-94-6 (2014)，p. 1-16
5. ウメルジャン・オスマン、中平勝子、鈴木俊哉、植村俊亮、三上喜貴：“ウイグル文字古文献デジタル化のためのグリフデザインの検討”，情報処理学会研究報告，査読無，DD-90-5 (2013)，p. 1-8
6. 川幡太一、鈴木俊哉、永崎研宣、下田正弘：“悉曇文字の国際標準化の動向”，情報処理学会研究報告，査読無，DD-90-7 (2013)，p. 1-4
7. 鈴木敦、鈴木俊哉：“A Survey of the Achievements of the Oracle Bone Digitization Projects and Prior Definitions of the Targets”，情報処理学会研究報告，査読無，DD-90-6 (2013)，p. 1-8
8. 鈴木俊哉：“HTML 文書へキャラクタデバイスの接続”，情報処理学会研究報告，査読無，DD-90-1 (2013)，p. 1-8
9. 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 回セミナー概要集 (2013)，査読無，p. 15-74
10. 鈴木俊哉：“文字分類方式の変更が字形に及ぼす影響”，情報処理学会研究報告，査読無，DD-86-3 (2012)，p. 1-6

〔学会発表〕（計 2 件）

1. 鈴木俊哉：“Toward to the Definition of Safe Character Set of Nushu in ISO/IEC 10646”，Japanese Association of Digital Humanities Conference 2014，査読有，2014 年 9 月 21 日，筑波大学
2. 鈴木俊哉：“『高麗大藏經異體字典』の画像分解”，情報処理学会デジタルドキュメント研究会第 94 回研究会ライト

ニングトーク，査読無，2014 年 7 月 24 日、広島大学東京オフィス
3. 鈴木敦、鈴木俊哉：“古漢字国際標準化の 10 年(2003-2012)”，文字研究会第 8 回研究会，査読無，2012 年 12 月 22 日，京都大学東京オフィス

〔図書〕（計１件）
飯島武次教授退官記念論集『中華文明の考古学』（同成社，2014，ISBN 9784886216588）、「甲骨文字研究の成果蓄積とデジタル化技術 － 近年の中国・台湾における動向を踏まえて」（p.101-111）、鈴木敦

〔その他〕
ホームページ等
本課題で作成されたデータについては
http://glyphsv.ipc.hiroshima-u.ac.jp/~mpsuzuki/OldHanzi/
で順次公開していく予定である。

標準化関連文書
1. 鈴木俊哉：“Feedback on Siddham proposal (WG2 N4294)”，ISO/IEC JTC1/SC2/WG2 N4361
2. 川幡太一、鈴木俊哉、永崎研宣、下田正弘：“Proposal to Encode Variants for Siddham Variants”，ISO/IEC JTC1/SC2/WG2 N4407
3. 鈴木俊哉：“Comments on Nushu scripts in N4484/PDAM1”，ISO/IEC JTC1/SC2/WG2 N4513
4. 鈴木俊哉：“Brief Summary of the Discussion about Shape-Based Separation of Siddham Vowel Sign U/UU”，ISO/IEC JTC1/SC2/WG2 N4557
5. 鈴木俊哉：“Comments on Nushu in ISO/IEC 10646:2014 PDAM2 (WG2 N4569)”，ISO/IEC JTC1/SC2/WG2 N4610

６．研究組織

(1)研究代表者
鈴木　俊哉（Shunya Suzuki）
広島大学・情報メディア教育研究センタ・助教
研究者番号：70311545

(2)研究分担者
鈴木　敦（Atsushi Suzuki）
茨城大学・人文学部・教授
研究者番号： 00272104

三上　喜貴（Yoshiki Mikami）
長岡技術科学大学・原子力安全系・教授
研究者番号： 70293264