Vol.2014-CH-103 No.5
2014/8/2

古漢字データベースの要件に関する試論

守岡 知彦[1]

概要：甲骨文字や楚簡資料等の先秦時代の漢字には現代漢字に共通する性質と異なる性質が存在する。これらの文字やテキストには解読中の部分が含まれることからその性質を記述する上で現代漢字に比べて学説依存性が高くなりがちである。一般的な漢籍資料に比べてテキストの量が少なく、楚簡の文字を推測する場合にも金文や甲骨文字、あるいは、説文解字等のさまざまな情報を併用しなければならないことも少なくない。こうしたややこしさを持つ文字の知識を記述する上での課題を挙げ、適切な電子化を行う上での要件について議論したい。

1. はじめに

漢字という文字システムを理解する上で、その歴史的変遷に関する知識は重要である。その意味で 19 世紀末に発見された甲骨文字資料と 20 世紀末に発見された楚簡資料は先秦時代の漢字の姿を知る上で重要な資料群であるといえる。もちろん、こうした文字学的興味以外にもそこに書かれたテキストの内容や言語、素材等は言語学、歴史学、文学、考古学等さまざまな分野における重要な資料といえ、その適切なデジタル化・機械可読化は人文情報学にとっても重要な課題だといえる。

甲骨文字や楚簡文字等の先秦時代の漢字は漢代以降の現代漢字の先祖の一つといえ、現代漢字と共通する性質を持ち対応する部品・文字も多数存在している半面、現代漢字とは異質な部分や現代では廃れてしまった部品・文字や対応がよく判らない部品・文字も多く存在する。これらの文字やテキストには解読中の部分が含まれることからその性質を記述する上で現代漢字に比べて学説依存性が高くなりがちである。一般的な漢籍資料に比べてテキストの量が少なく、楚簡の文字を推測する場合にも金文や甲骨文字、あるいは、説文解字等のさまざまな情報を併用しなければならないことも少なくない。

先秦時代の漢字の持つこうした性質は UCS [1] をはじめとする符号化文字モデルに基づく文字符号化・標準化を行う上でさまざまな困難を生じさせるといえる。UCS ではスクリプト（script, 用字系）毎に文字概念を抽象文字として符号化することを原則としているが、そのためには、まず先秦時代の漢字をどのようにスクリプトに分けるかが問題となる。しかしながら、文字数が多い漢字系文字を複数のスクリプトに分けると非常に広大なコード空間を消費することになるし、また、字体差が大きい先秦時代の漢字の抽象文字を符号化する場合、その例示字形にはあまり意味がなくなってしまい、対応する現代漢字で符号化するのと大差ないかもしれず、実質的に、同じ漢字を書体毎に重複して符号化するような結果になりかねない。一方、文字の解釈が明確でなく学説が分かれることも少なくないことから、文字概念が揺れている部分が多数あることになり、適切に抽象文字を定義するのも容易ではない。しかしながら、そうした研究中で学問上の論争がありそうな文字程論文で使われやすく、解釈が安定している部分は需要が少なくなりがちというジレンマがあり、符号化の経済性という観点で難しい問題をはらんでいるといえる。一方、文字システム的に漢字と異質な部分適切に扱うためには漢字とは別のスクリプトとしてそのモデルを適切に分析し、表現することも必要となってくるといえる。

本発表では、甲骨文字資料と楚簡資料を例に先秦時代の文字資料を適切に電子化する上での課題や要件について議論したい。

2. 先秦時代の書記系の特徴

漢字は古典中国語の形態素の多くが 1 音節であることを利用して、1 文字= 1 音節= 1 形態素となるようにした表語文字の一種であるといえる。ここで、多音節の形態素をどう表記するかが問題となるが、現代漢字ではそういう場合、1 文字= 1 音節であることを優先して、複数の文字で形態素を表現するようにしている。しかしながら、先秦時代には多音節の形態素を 1 文字で表したものがあり、これを「合文」と呼ぶ。合文の多くは複数の部品を組み合わせ

[1] 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

© 2014 Information Processing Society of Japan

たものであり、発音が推測できない場合、合文かそうでないかを判別することは困難だといえる。楚簡の場合、1文字がほぼ同じ大きさの仮想的な箱（仮想ボックス）の中に書かれ、仮想ボックス間に空白があるために、空白で区切られた仮想ボックス内に書かれたものを1文字と看做すことができる（図1）が、甲骨文字の場合、書き方の自由度が高く、文字の大きさにもばらつきがあるので、どこまでが1文字なのか判りづらいことがある（図2）。また、当時の発音は基本的に推測によらざるを得ないという問題もある。

図 1　楚簡の例

先秦時代の書記系の特徴としては記号の利用が挙げられる。一般的な古典中国語（漢文）の版本の場合、注を除いたテキストの本文中では特別な記号を用いず漢字だけを書き連ねるようになっているが、楚簡の場合、重文符号、文・語末記号、テキストの終りを示す記号が使われている。この内、重文符号は甲骨文字にも存在し、こうした記号の利用は古い時代の漢字テキストの特徴といえるのかもしれない。

重文符号（二）は文字の右下に付加される記号で、重文符号が付いた文字の文字全体あるいはその中に含まれる部品の繰り返しを示す。例えば、「夫$_{=}$」は「夫」の中の「大」を繰り返して「大夫」を示す。また、連続する複数の文字に付けて文字列の繰り返しを示す場合もある。例えば、「厲王生洹$_{=}$王$_{=}$即立」は「厲王生洹王、洹王即立」のように解釈できる。また、指定された文字が合文であることを示すために用いられた例もある（図1の1行2文字目の「㠯小子$_{=}$」）。また、これらが混在することもある。例えば、「㠯至于$_{=}$厲$_{=}$王$_{=}$大瘧于周」は「至于厲王、厲王大瘧于周」のように解釈できる。

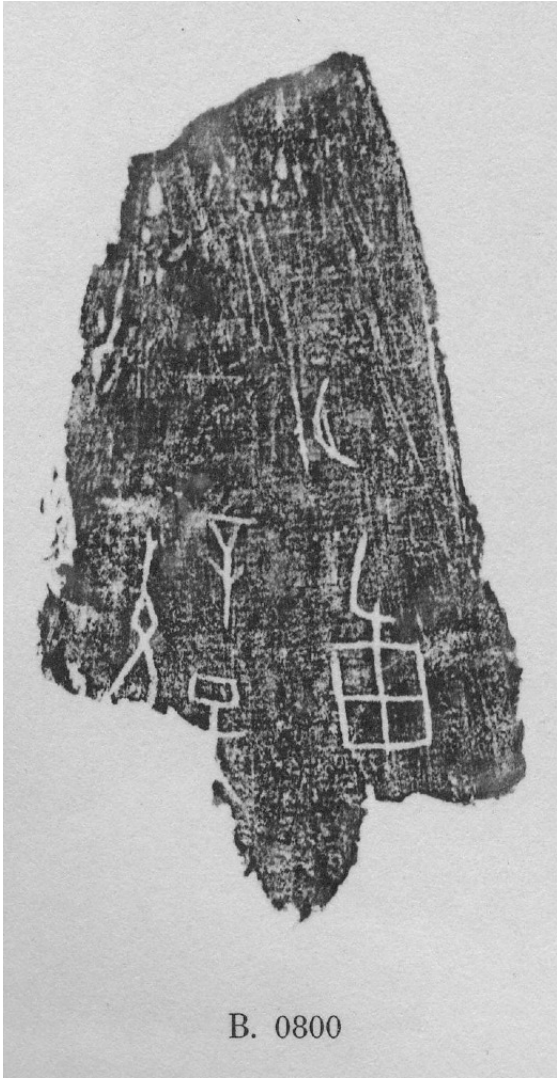

図 2　甲骨の拓本の例

文・語末記号は「、」に似た形をしており、重文符号と同様に文字の右下に付加される。この記号も用例毎に異なった使われ方がされており、未知の用法もあるかもしれないが、今の所、文の終りを示す場合と語（名詞、固有名詞等）の終りを示す場合とが知られている。

テキストの終りを示す記号は「し」に似た形をしており、あるまとまったテキストの終りを示す。

3.　何を何のためにデジタル化するのか?

先秦時代のテキストが書かれた甲骨や青銅器、竹簡・木簡等は貴重な文化遺産であり、もろく壊れやすいこともあり、保存のために実物へのアクセスが難しいことも少な

くない。そのため、デジタル代替物としての写真や拓本といった画像データは重要である。

検索可能性を考慮すれば、文字を符号化し、テキストを符号化文字列として電子化することが重要だといえる。単にプレインテキストにするだけでなく、文書構造に従ってマークアップしたり、画像上の位置情報と対応付けたり、アノテーションを付与したりといったことも考えられるが、こうしたことはテキスト自体のデジタル化を前提としたことであるといえ、そのためには文字の符号化が必要であるといえる。[*1]

先秦時代の漢字を符号化する場合、甲骨文字や楚系文字といった古い時代の漢字の字体そのものを整理して現代漢字とは別のスクリプトとして符号化するという方法と、文字や部品の現代漢字との対応関係に着目して対応する現代漢字（隷定字）を現代漢字と同じスクリプトとして符号化するという方法が考えられる。

前述のように、先秦時代の漢字には現代漢字と対応する文字や、文字としては対応するものがなかったり不明だったりしても部品としては対応するものがあり、部品の組合せ方（漢字構造）に着目して部品を現代漢字のものに置き換えることで機械的に現代漢字化（隷定）できることが多い。

読みやすさや検索のしやすさ、古典中国語のテキストとの比較のしやすさ等を考えた場合、隷定して現代漢字化した方が便利である。しかしながら、現代漢字への変換には解釈に依存する部分があり、また、現代漢字とは異質な部分を持つ文字体系を現代漢字からの視点でのみ解釈することの危険性も考えられる。

一方、甲骨文字や楚系文字を現代漢字とは別のスクリプトとして符号化しようとする場合、字体の揺れの大きさも問題であるし、また、現在研究中の部分が多く含まれ、文字の解釈の学説依存性が大きく、字体の包摂規準を定めて抽象文字を設定することが難しい場合が多々あるといえる。[*2] もちろん、解釈が安定していて、現代漢字との対応がはっきりしているものもあるが、こうした場合、逆に隷定して現代漢字で表しても問題が少なく、別スクリプトとして符号化することは実質的に書体の異なる漢字を符号化する結果につながりかねず、漢字の文字数の多さを鑑みればあまり望ましくないといえる。あるいは、なるべく解釈をさしはさまず、似た字形をまとめた字体や、あるいは、字形そのものを別字として符号化することも考えられるが、この場合、文字数が著しく増え、また、文字符号としての利便性を損ないかねない。字形を見たい場合、画像を見れば良い訳であり、画像を見ずに解読作業を行うことはできないのであるから、なんらかの抽象化を行わずに字形そのものを符号化するというのはナンセンスであるといえる。

甲骨文字や楚系文字を現代漢字とは別のスクリプトとして符号化するとして、それは何のために、そして、誰のために符号化するのかということを考えておくべきだろう。CHISE [2] のような自由に文字をその性質に基づいて定義可能で、その定義も頻繁に変更しても良いようなシステムを前提とする場合、その文字が使われた時代におそらくそうであっただろうと推測される性質に基づいて文字のオントロジーを構成し、文字オントロジー中の文字オブジェクトを用いてテキストを表現することはそれほど難しいことではない。この場合、もし新たな発見があったり間違いを見つけたら文字オントロジーを修正すれば良い訳である。しかしながら、UCS に提案し標準化する場合、一度 UCS に収録されたものを修正することは容易ではないため、こうした文字定義のリファクタリング作業は行えない。よって、スクリプトや文字概念が明確でないものを符号化するのは危険性が大きく、符号化（標準化）作業には大きな労力が必要になるといえるが、現代漢字程大きな需要がある訳ではなく、標準化によって得られる利益がその労力に見合うかどうかはなはだ疑問である。

そもそも、仮に先秦時代の漢字を UCS に収録した場合、誰がどのような目的で使うかということについて考えておく必要があるだろう。

仮に甲骨文字が UCS に収録され、パソコンやスマートフォン等で使えるようになれば、その視覚的形状の面白さから携帯絵文字のように使われることもあるだろうが、この場合、文字や文字体系のセマンティクスを真面目に定義する必要はなく、網羅性も要求されておらず、あまりシリアスな用途とはいえない。漢字辞書や国語の教材等で漢字の成立ち等を説明するために甲骨文字や金文等を出すといった用途も考えられるが、こうした用途では現代漢字との対応関係だけが要求されており、先秦時代の文字体系自体はどうでも良く、この意味では、携帯絵文字的な用途と大差ないかもしれない。

一方、先秦時代の文字資料を扱う研究者向けの用途として、先秦時代の文字資料のデータベースでの符号化手段としての用途と研究論文やネットでの議論等での利用が考えられる。データベースでの利用の場合、なるべく解釈をさしはさまない方が良いような場合には UCS は不向きであろうし、そうでない場合、現代漢字で隷定したものだけで良いかもしれない。研究論文等での利用の場合、過去の定説を参照するような用途もあるが、未解読のものや論争中の部分を扱うことも少なくないといえ、UCS で標準化し

*1 画像マークアップを行い、各文字の領域にその文字の性質に関する情報を付与することは考えられるが、これは結局、CHISE の Chaon モデル [2] に基づき、各字形、すなわち、文字のインスタンスを記述しているに等しい。また、実際には、字形情報の記述だけではなく、字体や字種といった文字のクラスに相当するものとの関係を書く必要もあり、これは文字を符号化する行為に他ならないといえる。

*2 UCS における甲骨文字に関する標準化動向を紹介記事として [3], [4] を挙げておく。

やすい部分だけでは不十分であり、逆に、標準化しにくい部分程需要があるといえる。微妙な差異を扱いたい場合、写真や拓本等の画像を扱わざるを得ず、また、それで問題ないといえる。

しかしながら、論文のアーカイブや書誌情報データベースや検索システムを考えた場合、論文等で使われた先秦時代の文字が UCS に収録されていないと問題を起こす場合がある。例えば、「釋甲骨文中的“𠂤”和“𠂤”」や「郭店楚簡「𠂤」構形試探」というように、論文の題名に先秦時代の文字を使った例がある。近年の楚簡研究の盛り上がりとともに、東洋学文献類目で必要とする数十個程度の外字の大半が楚簡文字とその隷定字となっており、中国学関連の学術情報サービスにとって無視できない影響を与えつつあると考えられる。書誌情報データベースは専門家だけが使うものではなく、図書館の司書や一般の利用者のことを考えれば、なるべく特殊な扱いは避けた方が良く普通に文字として扱えた方が良いといえる。また、XML 等との親和性を考えると、外字は望ましくなく、UCS に収録する方が望ましいといえる。ただ、例に挙げた論文の題名を見ても分かるように、微妙な字形差や解釈が不明だったり論争中だったりするものが使われることが少なくなく、適切に抽象文字を定義することが難しい例が少なくない。

いずれにしても、先秦時代の文字を符号化するのだとしても、現代の利用者がどのように文字を利用するのか、どういったコミュニティーでどのように利用されるのかといったことを分析する必要があるといえる。言い替えれば、先秦時代の文字を『現代の文字』としてどう位置付けるかという視点が重要なのではないかと思われる。

4. 書誌情報の中の先秦文字

先秦時代の文字に関する論文には、「甲骨文“𠂤”字考釋」のように、タイトル中で先秦時代の文字を使う例がある。京都大学人文科学研究所附属東アジア人文情報学研究センターが毎年刊行している東方学に関する研究論文や書籍等の目録である「東洋學文獻類目」(類目) では毎年十数から数十文字程度の外字が必要となっている (表 1) が、その多くは先秦時代の文字もしくはその隷定字 (である (表 2)。

	2001	2002	2003	2004	2005	2005 補	2006
種類	13	12	10	11	15	1	41
のべ	15	15	14	11	16	1	42

表 1 類目外字の出現数

次に、類目外字の再利用性について検討するために、各年版毎の新規外字の数と、その内、複数回利用された数と、別の年に再利用された数を調査し、表 3 にまとめた。このことから、誤って設けられたものを除いた、実際に使われた外字の、平均利用回数は 1.2 回であり、大部分の外字は

年度	甲骨	金文	楚簡	隷定	その他漢字	その他
2001	2	0	0	6	4	3
2002	0	1	1	10	3	0
2003	2	0	0	11	0	0
2004	0	2	0	7	1	0
2005	1	0	1	11	1	2
2005 補	0	0	0	1	0	0
2006	6	4	5	25	1	0
合計	11	7	7	70	10	5

表 2 字種毎の出現数

1 回限りの使用であることが判る。

	新規外字	複数回利用	複数年度利用
2001	13	2	1
2002	12	2	2
2003	8	2	0
2004	8	2	2
2005	13	1	0
2005 補	1	0	0
2006	38	0	—

表 3 新規外字の再利用度

5. 文字の符号化をどうするか

5.1 隷定字の場合

多くの隷定字は、漢字部品そのものは既に存在することが多いので、こうした場合、IDS (Ideographic Description Sequence) [1] 等で漢字構造を記述することで、代替表記を行ったり、外字管理をすることができるといえる。

しかしながら、「迷」のように、部品の一部 (この場合、「⺍」) が先秦時代の字体のまま現れる場合 (これを、ここでは『部分隷定字』と呼ぶことにする) がある。こうした場合、先秦時代の字体の部品の符号化ができないと、漢字構造を記述することができない。

5.2 甲骨文字の場合

「釋甲骨文中的“𠂤”和“𠂤”」といった細かな差異を問題にしたり、論争中の問題を扱った論文が少なくないため、一般的なカタログの論争の少ない部分だけを符号化したとしても、論文のタイトル等の電子化には不十分であると考えられる。

一方、部分隷定字の漢字構造記述のことを考えれば、基本部品の符号化には一定の価値があるのではないかと考えられる。また、甲骨文字自体の記述においても、『甲骨 IDS』(図 3) [5][6] のような漢字構造記述アプローチには一定の有用性があるといえ、そのためにも、有用であると考えられる。

5.3 金文・戦国期の文字の場合

個々の文字の符号化の問題以前に、これらをどのような

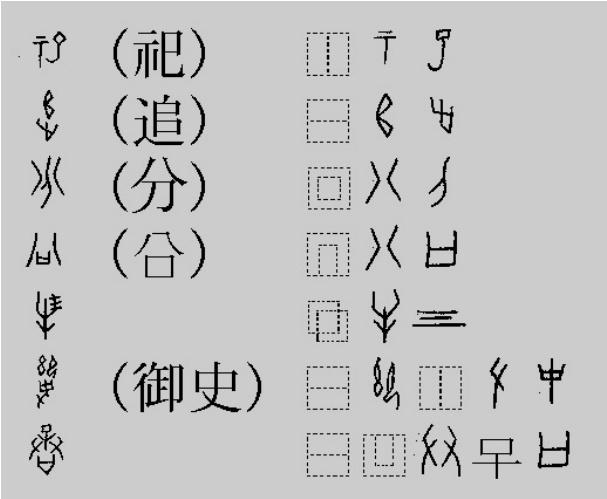

図 3　甲骨 IDS の例

スクリプトに分類するのか（あるいは、しないのか）という問題があるが、図書館情報処理の観点からいえば、ある程度まとまった数の論文が産出されるような研究分野・コミュニティが存在し（認知され）ているかという観点も重要なのではないかと考えられる。また、こうしたものを成り立たしめるものは、結局の所、まとまった数の出土資料等であるといえる。[*3]

そういう観点からいえば、今後、新たな発見によって、新たな字体の文字資料群が発見された時、それをまた新たなスクリプトとすべきかという問題が考えられるが、著者は、それによって研究分野が形成されるならば、それを新たなスクリプトとした方が実務上都合が良いだろうと考える。即ち、図書館情報処理の観点からいえば、必要なのは、現在、現実に論文のタイトル等で使われ書誌情報に記載されるような文字であり、それを司書が同定できるためにはある程度明確な研究分野が存在する必要があると思われるからである。

個々の文字の符号化に関しては、甲骨文字の場合と同様の問題があるといえる。よって、まず、基本部品の符号化を目指し、IDS 的な漢字構造情報の記述が可能となるような環境の実現が重要であると考えられる。

また、先秦時代の字体のまま使われる場合よりも隷定される場合の方が多いので、隷定字に対する配慮も重要であろう。

5.4　UCS に収録すべきもの

類目外字の大部分は再利用性が極めて低いが、中には例外的に繰り返し再利用される文字もある。この種の文字は、伝統的な符号化文字アプローチに基づき、UCS に収録することを検討する価値があるといえる。

*3　甲骨文字も殷墟の発見等で、まとまった数の出土資料が発見されたから甲骨文字として認知し得たといえる。

5.5　既に UCS に入っているもの

UCS の統合漢字拡張 C に「殷周金文集成引得」をソースとする漢字が 366 文字入っている [*4] が、これらの文字は金文の隷定字としての性格を持っているといえ、楚簡の隷定字とも重なる部分がある。

6.　CHISE でのアプローチ

6.1　文字の表現

先秦時代の漢字は、原則として、1 文字で 1 音節を表しその音節で表現される形態素を指示する表語文字であり、多くの文字が部品の組合せからなっており、異体字関係が存在する等といった現代漢字と同様の性質を持っており、文字を符号化したり文字に関する情報を記述する場合も現代漢字と同様に考えることができる。

但し、先秦時代の漢字の字体は楷書のように画数がはっきりせず、部品のデザインが楷書や明朝体といった現在一般的に使われる書体のものと大きく異なることから現在の一般的な漢字辞書で使われる康熙部首は使いづらい。こうしたことから、金文や楚簡、甲骨文字といった先秦時代の漢字に関する工具書は説文部首を使って分類・配列することが多い。文字情報のデータベースでは単一の配列法にこだわる必要はないが、説文部首の情報もあった方が無難であろう。[*5]

もちろん、検索の便利さを考えれば、CHISE 漢字構造情報データベースのような、IDS 等で部品の組合せ方の情報（漢字構造情報）を記述しておくことが望ましいといえる（図 3）。

現代漢字との対応関係が判っている場合、関係素性を使って現代漢字（隷定字）と先秦時代の文字との間にリンクを張ることが望ましいが、この時、どういうラベルを付けるかは問題である。CHISE 文字オントロジーでは対応する異なるスクリプトの文字オブジェクト間の関係素性名は、各ワードの先頭文字だけを大文字にしてハイフンでつなげた素性名を使うことにしており、例えば、甲骨文字から現代漢字には `<-Oracle-Bones`, 現代漢字から甲骨文字には `->Oracle-Bones` という関係素性を用いている。楚簡の文字の場合も、これを独立したスクリプトとみなして現代漢字とは別のオブジェクトを設けた場合、同様なスクリプト間関係素性を付与すれば良いといえる。

一方、部品や漢字構造に着目して隷定した場合、現代よく使われている字体とは異質な漢字になることは少なくないが、こうした隷定字は形式的には現代漢字と同一のスクリプトの文字であるといえ、隷定字と対応する一般的な漢字との関係にはスクリプト間関係素性を使うことができ

*4　[7] にその調査報告がある。
*5　説文部首ではなく甲骨文字の字形のみを抽象化して扱った分類体系として、1935 年に唐蘭が提案した「自然分類法」がある。[4] では自然分類法に基づく甲骨文字のデータベース化を提案している。

ない。

漢字辞書における伝統的な分類である「古文」(古字) はこのケースに近いものを含んでいると考えられる。実際、他により確かな情報源がない場合には説文解字の説をヒントにせざる得ないようである。CHISE では古文を関係素性 `<-ancient`, `->ancient` を用いて表しているので、例えば、楚簡の隷定字に対してもこの関係素性を使うことが考えられるが、機械可読性を重視するという観点では、指示対象のドメインだけでなく、研究史的な文脈やコミュニティー等の観点に基づいたドメインを適切に表現することが望ましいといえ、伝統的分類の中に近年の研究成果の知見を何のアノテーションもなしに混ぜてしまうのは問題かもしれない。こうしたことから、現在の所、楚簡の隷定字と一般的な漢字との対応関係には、楚簡を表すドメイン 'chu' を付け、`<-ancient@chu` のように表すことにしている。

6.2 合文の表現

合文の問題、あるいは、どこまでが 1 文字かはっきりしないケースは文字を符号化したり文字の性質を記述しようとする場合には問題であるが、実の所、テキストの解釈を行う上では割とどうでも良い問題だといえる。つまり、どういう形態素を表現しているかが分かるのであれば、別にそれが 1 文字だろうが複数文字であろうがどうでも良いし、また、正確な発音を直接知る術がない以上、それが本当に 1 音節だったのか多音節だったのかも原理的にはっきりとは知り得ないといえる。

文字資料のデータベースとしては、事実と解釈が適切に区別され、複数の解釈が両立可能な枠組になっていることが望ましく、また、研究者にとってあまり重要でない問題に大きな労力を割くのはあまり望ましくないといえる。合文の問題を適切に扱うためには、結局の所、文字レイヤーだけではなく形態素レイヤーと密接に連携した形で扱うのが望ましいといえる。

現在、CHISE 文字オントロジーでは MeCab 用古典中国語形態素コーパス [8] を Linked Open Data にしたもの (古典中国語形態素 LOD) [9] を文字用例としてリンクしているが、これと同様な形で先秦時代の文字資料に対しても形態素 LOD を実現することで、文字オブジェクトと形態素オブジェクトとの相互リンクを実現することが可能である。先秦時代の文字資料の形態素コーパス化は古典中国語の場合以上に間違いやすく形態素列としての記述の揺れが生じやすいと考えられるため、形態素 LOD 化することで、同一の文に対する異なる形態素列や異なる品詞体系等を共存可能にすることも重要であると考えられる。

6.3 重複符号の扱い

2 節で述べたように、重複符号の振舞には幾つかのパターンがあり、正しい解釈かどうかを判定するには文法的・意味的な妥当性を調べる必要があり、機械的に正しい解釈を得るのは容易ではないといえる。重複符号の展開結果は 1 つの形態素に収まる場合もあり、その場合、形態素の表記として扱えば良いが、形態素や句・文の境界をまたがる場合もあり、形態素解析の前にあらかじめ重複符号を展開する処理が必要となり、形態素コーパスの記述にも工夫が必要となる。

竹簡が不鮮明な場合、重複符号と文・語末記号「、」の区別が付きにくい場合があるので、重複符号を展開した形でコーパスを作った場合、リカバリー処理が難しくなる。よって、仮に重複符号を展開するとしても、元表記との対応が付くようにしておき、展開した形のコーパスから元表記のコーパスに戻れるようにすべきであるといえる。

6.4 上古音の情報

楚簡のテキスト中で文字を解釈する時、用例が少なくその性質がよくわかっていない場合を中心に、対象となる文字を別のよく知られた音の似た文字を表しているものと仮定して推論する場合がある。この時の音の類似性は「上古音」と呼ばれる周〜漢代頃の音韻体系を復元したものをベースに考えることになるため、この情報を機械可読化しておくことが重要であるといえる。

6.5 文字の典拠情報

先秦時代の漢字の場合、現代漢字以上に学説依存性が高いといえ、また、学説が新たな学説によって書き換えられる頻度も高いといえる。しかしながら、こうした文字の解釈に関する学説は、多くの場合、論文や書籍、ブログや Web 掲示板等で公表されるので、対象となる文字オブジェクトやその素性に、メタデータ素性等を用いて、文献を紐付けておくことで、より具体的な典拠情報を知る手がかりにしたり、典拠の種類に応じた異体字関係の制御を行うことが考えられる。

類目外字の場合、出典となる論文や書籍が存在し、類目外字がどこで使われたかを容易に把握できるので、文字の典拠に当たることが比較的容易である。CHISE の文字情報にアクセスするための Web UI である CHISE-Wiki [10][11] では、現在、東洋学文献類目のタイトルでの用例情報を載せるようにしており、ある文字がどういう文献で使われたかが分かるようになっている。これは類目外字だけでなく、UCS に収録された漢字に対しても行っており、ある漢字がどういう分野で使われているかを知るためのヒントになりうる。

一方、京大人文研所蔵甲骨文字の場合、京大人文研所蔵甲骨の拓本画像へのリンクを設けており文字の出典情報は得られるが、研究論文等との紐付けはできていない。

© 2014 Information Processing Society of Japan

また、いずれにしても、現状では、文字の典拠情報はあっても、文字素性単位での研究論文等との紐付けができていないが、先秦時代の漢字のより適切な文字処理のためには、関係素性をはじめとする文字の性質に関わる素性単位に典拠情報となる文献との関係を機械可読な形で記述することが重要であるといえる。

7. おわりに

甲骨文字資料と楚簡資料を例に先秦時代の文字資料を適切に電子化する上での課題や要件について議論した。

先秦時代の文字資料を適切に記述するには、文字単位の情報だけではなく形態素コーパス・辞書を組み合わせた Linked Data 的構成を採ることが望ましいと考えられる。こうした文字レイヤーと形態素レイヤー等の複数レイヤーからなるデータベース化の試みは古代エジプト神官文字のデータベース [12][13] やアマルナ文書の電子化 [14] といった漢字以外の古代文字資料の電子化においても行われていることであり、古代文字のデータベース化において必要な用件といえるのかもしれない。また、各情報に対してその根拠や典拠等を述べた研究論文を紐付けておくことが望ましいといえる。

謝辞

本論文の楚簡に関する記述は京都大学人文科学研究所共同研究班「清華大学蔵戦国竹書を読む—中国古代の基礎史料」で学んだことを私なりに解釈した結果である。班長の浅原達郎先生と班員諸氏に感謝する。特に、浅原先生は全くの初心者である私の愚かしい質問に対していつも丁寧にご教示して頂いた。こうした、親切で丁寧な説明にも関わらず、おそらく私の誤解に基づく間違いがあるかもしれないが、これは無論、全て私の責任であることはいうまでもない。

参考文献

[1] International Organization for Standardization (ISO): *Information technology — Universal Multiple-Octet Coded Character Set (UCS)* (2011). ISO/IEC 10646:2011.

[2] Morioka, T.: CHISE: Character Processing based on Character Ontology, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No. 4938, pp. 148–162 (2008).

[3] 小形克宏：UCS における甲骨文字収録の意義と問題点，東洋学へのコンピューター利用第 18 回研究セミナー，pp. 151–173 (2007).

[4] 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ー利用第 24 回研究セミナー，pp. 15–74 (2013).

[5] 守岡知彦：CHISE に基づく甲骨文字資料の電子化について，人文科学とコンピュータシンポジウム論文集デジタルアーカイブ —デジタルアーカイブと時空間の視点—，15, 情報処理学会，情報処理学会，pp. 115–122 (2007).

[6] 守岡知彦：甲骨文字処理にまつわるエトセトラ，東洋学へのコンピューター利用第 19 回研究セミナー (2008).

[7] 鈴木俊哉：統合漢字に申請された「殷周金文集成引得」図形文字の調査，情処研報，Vol. 2009-DD-72, No. 7, pp. 1–8 (2009).

[8] 山崎直樹，守岡知彦，安岡孝一：古典中国語形態素解析のための品詞体系再構築，人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集，情報処理学会シンポジウムシリーズ，Vol. 2012, No. 7, 情報処理学会，情報処理学会，pp. 39–46 (2012).

[9] 守岡知彦：古典中国語形態素コーパスの Linked Data 化の試み，人文科学とコンピュータシンポジウム「じんもんこん 2013」論文集，情報処理学会シンポジウムシリーズ，Vol. 2013, No. 4, 情報処理学会，情報処理学会，pp. 187–194 (2013).

[10] 守岡知彦：CHISE のセマンティック Wiki 化の試み，情処研報， Vol. 2010-CH-87, No. 8, pp. 1–8 (2010).

[11] 守岡知彦：Wiki 的手法に基づく構造化データの編集について，人文科学とコンピュータシンポジウム論文集—人文工学の可能性〜異分野融合による「実質化」の方法〜，情報処理学会シンポジウムシリーズ，Vol. 2010, No. 15, 情報処理学会，情報処理学会，pp. 33–40 (2010).

[12] 永井正勝：古代エジプト神官文字の画像を利用した字形データベースについて：中エジプト語の資料を中心とした事例紹介，人文科学とコンピュータシンポジウム論文集：「デジタル・アーカイブ」再考—いま改めて問う記録・保存・活用の技術—，情報処理学会シンポジウムシリーズ，Vol. 2011, No. 8, 情報処理学会，情報処理学会，pp. 389–394 (2011).

[13] 永井正勝，和氣愛仁：古代エジプト神官文字写本を対象とした言語情報表示システムの試作，人文科学とコンピュータシンポジウム論文集 2012，情報処理学会シンポジウムシリーズ，Vol. 2012, No. 7, 情報処理学会，情報処理学会，pp. 225–230 (2012).

[14] 高橋洋成：アマルナ文書の電子化 —文字研究・言語研究を目指して—，情処研報，Vol. 2013-CH-99, No. 6, pp. 1–7 (2013).

© 2014 Information Processing Society of Japan