京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

京都大学
KYOTO UNIVERSITY

Title	古典中國語テキストの知識處理について
Author(s)	守岡, 知彦
Citation	東方學報 (2010), 85: 556-578
Issue Date	2010-03-25
URL	http://dx.doi.org/10.14989/131773
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

Kyoto University

東方學報 京都第 85 册(2010)： 578－556 頁

古典中國語テキストの知識處理について

守 岡 知 彦

1 はじめに

近年，計算機の處理速度や記憶容量の增大によって，かつては處理時間や記憶容量等の制約であまり現實的でなかったり難しかったような處理，例えば，自然言語處理やパターン認識，物理計算，大規模なテキストデータベースに對する複雜な處理といったことが次第に比較的安價に利用できるようになってきた。消費電力の問題から，計算機アーキテクチャは次第に竝列化の方向に進みつつあり，こうした現狀は從來難しかったこれらの分野の發展をさらに後押しすると考えられる。

一方，電子化の進展とともに，畫像やプレイン・テキスト等の『淺い』構造化を行ったデータは急速に增大してきているが，精緻にマークアップするといった『深い』構造化を行ったデータの蓄積はそれほど進んでいない。というのも，現狀では古典中國語のためのコーパスや各種データがあまり蓄積されておらず，現代語では利用可能な文字認識や自然言語處理等のさまざまなツールが利用できないために，作業の多くを專ら人手に負わざるを得ないからだと考えられる。

また，漢字文獻の場合，いわゆる異體字の問題，すなわち，字形や字體といった文字の視覺的表現（以下では，これを『グリフ』と呼ぶことにする）と意味との對應關係の問題を適切に扱うことが重要となってくるが，古典文獻の場合，この問題は解釋に關わる問題であるといえる。しかしながら，計算機にとっては何らかの『正規形』をベースにそこから派生する形で『表現型』（見掛け）に關わる情報が生成するというモデルに從って扱うのが容易である。テキストデータにとっての基礎となる UCS のような汎用文字符號もまたこうしたモデルに從って構成されている。この結果，現狀では畫像とテキストデータを併用し，別々にデータ化するのが樂だといえる。しかしながら，その結果，テキストの視覺的な情報・構造と論理構造や意味的な情報がバラバラに扱われることとなり，本來，兩者の對應關係として表されていたような情報を處理することが困難となっている。

こうした狀況を改善する上で重要なことのひとつは，現實的な自然言語處理技術の利用であり，もうひとつは，視覺的情報と論理構造や意味的な情報の對應關係を扱えるような

多面的な處理技術の確立であると考える。ここでは，文法コーパス，グリフコーパス，文字オントロジーというものをキーに，こうした問題を解決するための多面的漢字知識處理技術について概說したい。

2 『データを生み出すデータ』の重要性

2.1 『賢い』データを蓄積することの難しさ

計算機とインターネットの普及によって，我々は資料を電子化することの福音を知ることとなった。そして，同時に，電子化されたデータとの戰いが始まることとなった。なんらかのアプリケーションソフトでデータを入力する事自體は必ずしも技術的に困難なことではない。しかしながら，データを校訂し，メンテナンスし續けること，多量のデータを適切に扱うことには勞力と，そして，フレームワークとワークフローの適切な設計が必要である。そして，フレームワークの設計のためには對象領域に關する適切なモデル化が必要となる。ここに人文情報學的研究の必要性が生じる譯である。しかしながら，こうした意味での人文情報學的研究の蓄積は必ずしも多いとはいえない[1)]。そして，情報學的研究課題としてみた場合，人文學的對象はなかなかに手ごわい，難しい課題であることが少なくない。[13] だからこそ，大變興味深い對象であるということがいえるのだが，資料を電子化すればすぐに福音を得ることができるはずだという計算機に對する素朴なイメージと，現實の情報工學的な制約の間には大きな溝があることが少なくない。

著者自身もこれまで幾つかの人文系情報サービスに關わるような事業や研究プロジェクトに關わってきた。また，主に，個人的な關心のもと，CHISE プロジェクトにおいて文字オントロジーおよびその處理系の開發に取り組んできた。[3][9] こうした中で，否應なしにデータと人間，その勞力・コストの問題，責任や評價の問題，メンテナンス體制の問題，事業と研究の問題，プロジェクトと繼續の問題，マネージメントの問題，心の問題，等々のさまざまな問題の存在に氣づかされることとなった。こうした問題の全てをここで取り上げることはできないし，また，人文系情報サービスに關わる問題は更に多肢に渡る問題でもある。例えば，[4] で永崎氏は人文系デジタル・アーカイブにおけるステイクホルダーに關して論じているが，ここで述べられているように，本來，多數の當事者の共同作業によってなされる問題であるが，資料のデジタル化においてある種の歪が生じているという風に考えるべきなのかも知れない。また，このような見取圖がないまま，プロ

1）計算機を使って，何かを作ったり，計ったりするという研究に比べて。

ジェクト型で見切り發車して進んで行くことが多いことも問題の一因かも知れない。そして，人文情報學分野一般の問題ではあるが，その全貌を把握できる人材が少ない，ないしは，いないということが問題の解決を難しくしているかも知れない。そして，著者も極めて偏った見方でしか問題を見ることができない者の一人である。

しかしながら，この問題の幾つかの（重要な）要素は，人文系情報サービスに關わる人員・勞力の效率的運用の問題という風に考えることができ，ある種の資源配分問題に歸着するといえるかも知れず，ある意味，極めてソフトウェア工學的な問題と看做せるかも知れない。

ただ，世の中の一般的な情報サービスに比べて，人文系情報サービスではしばしば投入可能な人員に限りがあること（單純作業のための人員は比較的確保が容易であるが，專門的な知識や技能を要する人員の確保は難しい。これはある意味當り前なことであるが，人文系情報サービスの構築作業では高度なサービスを提供しようと思えば思う程，對象領域に關する知識が必要となるし，また，情報學研究者・技術者との深いレベルでの對話が必要となるが，こうしたプロジェクトチームを確保するのは容易ではない），問題が特殊であるために一般的なソフトウェアモジュールや方法論が使えないことがあること，長期に渡る資料（情報）の蓄積があるために，レガシー問題に直面することがあること，などといった點から，一般的なソリューションを適用しづらい場合もあるかも知れない。

2.2　ツールチェーン

一般に，ソフトウェア開發には開發環境の存在が必要である。例えば，C 言語による開發を行うためには，C コンパイラの他，libc をはじめとする基本ライブラリ，リンカー，アセンブラー等が必要であり，このためには，ABI（Application Binary Interface）の定義や各モジュール間のインターフェイスの定義も必要である。また，エディターやデバッガー等も必要であるし，各種ライブラリーや統合環境もあると便利である。こうした開發に關わるプログラムの集合體を『ツールチェーン』（tool chain）と呼ぶ。

人文系情報システムの開發の難しさの要因の幾つかは，こうしたツールチェーンの缺如であると考えられる。例えば，漢籍の場合，古くから漢字との戰いの問題があり，近年では UCS統合漢字擴張 B（Ext-B）［1］が通るかどうかという問題があって，この問題のために使いたいシステムが使えなかったり，改造をする必要があったりした。こうした問題は徐々に改善されつつあるが，現時點でも問題に遭遇することは少なくない。また，人文系情報システムが扱うべき對象は，しばしば，言語や意味の問題に關わっている。よって，形態素解析器や構文・係り受け解析器といった自然言語處理に關わるツールが利用できると良いのであるが，古典語やマイナーな言語の場合，處理系の種類が極めて少なかった

り，存在しなかったりする。日本語や英語といったメジャーな現代語の場合，形態素解析や構文解析のためのある程度實用的に利用可能なツールが整備されており，全文檢索といったカジュアルな用途でも利用されるようになってきている。例えば，古典中國語で自然言語處理に關わるツールが利用できないということは，現代日本語だと普通にできることが古典中國語ではできないということを意味する譯である。こうした問題を解決するためには，自然言語處理に關わるツールやデータの整備が重要である。こうしたことは以前は容易ではなかったが，少なくとも形態素解析器に關しては MeCab のような言語に依存しない處理系が登場しており，辭書や文法コーパスを用意することで古典中國語の形態素解析器を比較的容易に實現することができるようになった。［10］より上位のレイヤーに關しては言語依存性が高いが，古典中國語のような多くの文獻・資料を有する言語の場合，構文解析器や意味處理のためのオントロジー等の整備も行うべきだろう。

2.3 基盤データ

前節で述べたツールチェーンの問題はデータに關してもいえる。紙の時代から工具書は研究のためのツールチェーンを構成していたといえ，電子化された情報リソースもまた研究のためのツールチェーンを構成することが期待される。しかしながら，電子化された工具書は必ずしも期待を滿たしているとは言い難い面がある[2)]。

電子化されたデータが存在すれば，とりあえずそれを檢索したり分析したりすることはできる。紙の情報と同じものが電子化されていれば，人間が見るには十分であるといえる。しかしながら，別の情報システムのための基盤となる情報リソースとしては單に人間が讀めるだけでは不十分であり，機械にとって理解可能な形式化がなされている必要がある。

こうしたことから，電子テキストのマークアップ化が行われたりオントロジーが作られるようになった譯であるが，こうしたデータの作成には多かれ少なかれ人手がかかり，詳細に形式化しようとすればする程勞力がかかる。よって，明確な目的・目標を定め，それを實現するための必要最低限の形式化を行った方が妥當であるといえる。

しかしながら，そうすると今度は他の目的への轉用がしにくいものとなりがちである。それでは結局，人間が見ることしか考えていない電子化と結果的にさほど變わらないことになってしまうといえる。實際，費用對效果を考えた結果，畫像とテキストのような，『淺い』形式化しか行わない例も增えてきたように思う。これはこれでひとつの見識ではあり，將來，パターン認識技術や自然言語處理技術が向上すればこうした情報も基盤データ

2）例えば，文字の場合［9］。

として活用できる時が來ると思われるし，そうした研究のための實驗用データとしても有用であるといえるが，現時點では，より高度な情報サービスを行うためにはそれなり形式化やメタデータの付加が必要であるといえる。

問題は，現時點での目的のために勞力をかけて詳細なマークアップや輕量でないオントロジーといった『深い』形式化を行う場合である。もし，多大な勞力をかけて『深い』形式化を行ったとしても，結果的に，『淺い』形式化と同程度のことしかできないとしたら，費用對效果の面で難があるといえる。實際には，『深い』形式化をすればなにがしか『淺い』形式化以上のメリットが得られるものではあるが，基盤データという觀點で考えた場合，データの適用範圍が廣くなければ，いかに目的とする對象での高度な檢索サービスが利用できたとしても，そこから抽出できる知識や應用範圍は『淺い』形式化の場合と同程度であるという風に考えることができる。言い替えれば，そのデータから構成し得る情報サービスの種類の數が同程度であるということである。

無論，これはユーザビリティ等の個別の情報サービスにおける質の問題を無視した議論である。ここで議論しているデータの適用範圍の廣さは，基盤データとしての有用性に關わる要素であり，現實の情報システムでは，こうした基盤データに關わる要素の他に，『アプリケーション固有データ』とでもいうべき，個別の情報システム固有のデータも必要だといえる。基盤データの役割はアプリケーション固有データ作成の勞力を下げるためにあり，全ての情報システムの集合を考えた時に，各情報システムで必要とするデータを作るための勞力の總和を下げ得るような共通知識が基盤データの役割といえる。

『深い』形式化には多くの人的コストがかかるので，『深い』形式化を行うのはなるべく基盤データに限るべきだといえる。無論，他のデータによって機械的に生成・導出可能な場合，アプリケーション固有データが結果的に『深い』形式化されたものとなるのは妥當である。しかしながら，他のデータによって機械的に生成・導出可能な情報の場合，人手で入力するのは費用對效果の點で妥當ではないといえる。

問題は，現時點では，基盤データの整備が進んでいないために，理論的に可能であったとしても，今すぐ現實に行うことはできないことが多々あるということである。このようなことから，從來，今すぐ使えるツールを使ってとりあえず電子化するということが行われ，結果的に，基盤データやそれによって構成されるべきツールチェーンの整備が進んで來なかったきらいがあるように思われる。

基盤データとしては，電子テキストの場合，文字處理に關わるもの，自然言語處理に關わるもの，意味處理に關わるものなどが考えられるが，古典中國語の場合，文字處理に關わるものを除けば，網羅的・體系的なデータ整備はあまり進んでいないといえる。こうした基盤データは互いに關連しており，相互變換・檢證のためのツール整備も含め，複數の

基盤データからなるツールチェーンを構成することが望ましいといえる。

2.4 データベースのリファクタリング

人文系資料の電子化を行う上での困難の一つは，仕様を明確に定義しづらいことだといえる。さまざまな暗默知が混じった對象分野の『常識』を分析して，明確に形式化された仕様を定義すること自體，人文情報學の重要な研究テーマになることが少なくないといえ，そのためには，長期に渡る參與觀察や學際的共同研究を要する場合も多いといえる。

一方で，關係データベースに基づく今日の一般的なデータベース・システムはスキーマの變更が簡單ではないという問題點がある。こうしたことを鑑みれば，データベースのリファクタリングを技術的に容易にすることで，人文情報系データベースにおける技術的な問題の幾つかを解決できるのではないかと考えられる。

このような問題は人文科學に限ったことではなく，自然科學系の研究においても，實驗データや測定データをモデルや假定・研究者の視點等に沿って評價・分析・整理・分類し，新たな假說を檢討するといったことを試行錯誤しながら行う場合があり，このような用途に對して **DREAM モデル**［11］［12］が提案されている。DREAM モデルは未整理の一次データに對して，それを元に研究者が作成した二次データを試行錯誤しながら作成・管理することを目的としたものであり，一次データを格納する「基本データベース」に對して二次データを表現するメタデータを張り付けて行くようなものとなっている（この二次データは「導出データベース」に格納される）。

また，全ての情報を狀況論的に扱うアプローチも有り得，著者は，文字に對して，CHISE の文字モデルである **Chaon モデル**［3］［9］を提案・實裝している。また，文字以外を含むオブジェクト一般を對象としたプロトタイプ方式のオブジェクト指向データベース **Concord**［8］を提案・實裝している。

データベースのリファクタリングを行うためには，テストケース等を用いることで意圖せざるバグの混入を防止することも不可缺である。また，データベースの形式・構造を變更する際には，データベースの等價變換を可能にする（ちゃんと變換できたことが機械的に確認できる）仕組みが必要であるといえる。

後者に關しては，例えば，變換プログラムと逆變換プログラムを對で作り，變換プログラムで變換されたデータを逆變換プログラムで逆變換して，その結果が元のデータと同じになった時に變換が成功したと看做すというような方法で實現可能である。また，できれば，この變換プログラムと逆變換プログラムの對とその時のデータベースのスナップショットを版管理することが望ましい。スキーマを頻繁に擴張するためには，前述の DREAM モデルや Concord，あるいは，RDF データベースのような，關係データベース

（關係モデル）以外の方式の利用を檢討したり，そのためのシステム開發を行うべきかも知れない。また，データベースの意味論的な檢證のために，オントロジーを利用するという研究も重要であるといえる。

3　形態素解析器の利用

日本語や中國語といったメジャーな現代語においては，近年，自然言語處理に關するツールの整備が進んできている。特に形態素解析に關しては精度や速度の點で安價な PC においても實用的に利用できるレベルに達しており，全文檢索やデータマイニング，テキスト處理等を支える重要な基盤のひとつとなってきている。しかしながら，文法コーパスの整備が進んでいない古典語やマイナー言語ではこうした利益を享受することが困難である。このことはこうした言語ではメジャーな現代語と同水準の檢索やテキスト處理サービスを享受することができないということを意味している。そして，古典中國語もまたこうした言語のひとつであるといえる。

こうした狀況を打開するには，古典中國語のための良質なタグ付き文法コーパスを整備し，形態素解析や構文解析といった自然言語解析のための基盤技術を整備していくことが望ましいのであるが，現狀では，０からツール，辭書，文法コーパスを開發するのはあまり現實的ではないといえ，既存の利用できそうなものはなるべく利用してなんらかのプロトタイプを實現し，それを元に少しずつ辭書や文法コーパスを改良していくような漸進的アプローチを採るのが良いと考えられる。ツールや辭書，コーパスは互いに絡み合っており，また，闇雲にデータを增やせば良いというものではなく，『次元の罠』という言葉に象徵されるように，データを增やしたことによってかえって認識率が下がるということもあり得る。無意味な勞力を避けるためにも，十分に處理を意識したデータ整備が望まれるといえ，そのためにも實際に動くプロトタイプを用意することは有效であるといえる。

そこで，著者は MeCab［２］と IPA 辭書という現代日本語用の形態素解析器をベースに，極めて少ない勞力で，古典中國語のための形態素解析器のプロトタイプの實現を試みた。

3.１　MeCab とは

MeCab［２］は工藤拓氏によって開發されている形態素解析エンジンで[3)]，オープン

3）京都大學情報學研究科－日本電信電話株式會社コミュニケーション科學基礎研究所共同研究ユニットプロジェクトを通じて開發された。

ソース・ソフトウェアとして公開されている[4]。MeCab は言語，辭書，コーパスに依存しない汎用的な設計がなされており，辭書，コーパス，品詞體系等を用意することで現代日本語以外の言語でもサポート可能な構造になっている。また，UTF-8 をサポートしており，UCS統合漢字［1］が利用可能である[5]。こうした特徴から，古典中國語のための形態素解析エンジンとして有望であるといえる。

MeCab で日本語の形態素解析をするためには，現在，IPA コーパスに基づく「IPA 辭書」と京都コーパスに基づく「Juman 辭書」が公開されており，この內，前者が推奬されている。

3.2 必要となるデータ

MeCab を利用するためには少なくとも辭書が必要である。MeCab は少なくとも辭書があれば形態素解析を行うことができるが，學習用コーパスからパラメータ推定を行うことで，連接コストを考慮した解析が可能となっている。

このパラメータ推定を行うには，

1．Seed 辭書

2．學習用コーパス

の 2 つが必要である。

Seed 辭書は利用者が用意する辭書であり，これを學習用コーパスとともに處理することで，連接コストの情報が付與された辭書が作成される。

3.2.1 辭書

MeCab の辭書はコンマ區切り（Comma Separated Value；CSV）形式になっている。最初の 4 カラム目：

1．表層形（單語そのもの）

2．左連接狀態番號

3．右連接狀態番號

4．コスト

は必須項目であり，2 〜 4 番目はコーパスからの學習によって自動生成される項目であり，Seed 辭書では使われないので 0 とすることになっている。

5 番目以降の欄は「素性」と呼ばれる項目で，利用者は好きなだけ素性を付與すること

4）GPL，LGPL，または，BSD ライセンスに從って使用，再配布することができる。

5）BMP の範圍は完全に對應しており，Ext-B も（少なくとも入力文字列としては）利用可能なようである。

ができるようになっている。品詞，活用，讀み，發音といった單語に關する情報はこの素性を用いて記述することができる。この特徴により，MeCab は言語，辭書，コーパス獨立性を得ているといえる。

3.2.2 學習用コーパス

學習用コーパスは，タブとコンマで區切られた複數の行と EOS のみの行で一文を表したものを連ねたものである（圖１）。

これは，MeCab のデフォルトでの出力と同一の形式であり，MeCab の出力を使って容易にコーパスを編集できるように設計されている。

```
智者	n,名詞,人,*,*,*,智,智者,チシャ,*
不	v,副詞,否定,否定,*,*,不,ず,ズ,*
惑	v,動詞,思惟,認識,*,*,惑,惑う,マドウ,五段・ワ行促音便
EOS
不	v,副詞,否定,否定,*,*,不,ず,ズ,*
遠	v,形容詞,*,*,*,*,遠,遠し,トオシ,形容詞・アウオ段
千	n,数詞,数,*,*,*,千,千,セン,*
里	n,量詞,単位,長さ,*,*,里,里,リ,*
EOS
無	v,動詞,否定,存在,*,*,無,無し,ナシ,形容詞・アウオ段
惻隠	n,名詞,一般,*,*,*,惻隠,惻隠,ソクイン,*
之	p,+p+,*,*,*,*,之,之,ノ,*
心	n,名詞,一般,*,*,*,心,心,ココロ,*
非	v,形容詞,否定,否定,*,*,非,非ず,アラズ,*
人	n,名詞,人,*,*,*,人,人,ヒト,*
也	p,+p.,語気,陳述,*,*,也,也,ナリ,*
EOS
夫	p,p+,語気,発語,*,*,夫,夫れ,ソレ,*
礼	n,名詞,一般,*,*,*,禮,礼,レイ,*
者	p,+p,+n,*,*,*,,者,モノ,*
自	n,代詞,人,一人称,単数,*,自,自ら,ミズカラ,*
卑	v,動詞,謙敬,謙譲,*,*,,卑くす,ヒククス,*
而	p,接続詞,*,*,*,*,而,而,*,*
尊	v,動詞,謙敬,表敬,*,*,尊,尊ぶ,タットブ,五段・バ行
人	n,名詞,人,*,*,*,人,人,ヒト,*
EOS
```

圖１　學習用コーパスの例

3.3 辭書の作成

3.3.1 IPA 辭書からの漢語の抽出

IPA コーパスは現代日本語のコーパスであり，當然のことながら，IPA 辭書を用いたのでは古典中國語の形態素解析はできない。しかしながら，現代日本語には古典中國語由來の語彙が多數含まれ，そうした語彙を切り出すことにより，古典中國語の辭書をでっちあげることができると考えられる。

さて，どういったものが古典中國語の語彙となり得るかであるが，漢字以外の文字を含む語は明らかに古典中國語の語彙とはなり得ないといえる。また，品詞の規則的對應が存在することも必要であり，助詞や助動詞などは機械的に變換できないといえる。

動詞および（文語）形容詞は，古典中國語においても，それぞれ，動詞および形容詞となるものが多いと考えられ，語幹を取り出すことにより，機械的に變換可能であるといえる。

同樣に，一般名詞は古典中國語においても名詞になるものが多いと假定できる。また，サ變名詞は動詞になるものが多いと假定できる。ただ，名詞の場合，古典中國語において句や文を構成していたものが日本語において名詞化したものが多數あり（例：非常，不祥），こうしたものを除去する必要がある。

こうしたことを鑑みて，次の品詞を持つ語彙を機械的に變換することにした（表1）。

表1　機械的に變換可能な品詞

IPA 辭書	古典中國語	變換語彙數
名詞（一般）	名詞（一般）	50577
名詞（副詞可能）	名詞（副詞可能）	801
名詞（數）	名詞（數）	26
名詞（サ變接續，１文字）	動詞	11030
動詞（基本形）	動詞	3991
形容詞（文語基本形）	形容詞	272

3.3.2 追加する必要のある語彙

IPA 辭書から機械的に變換した語彙だけでは，副詞や助動詞，助詞，前置詞，接續詞等の文法的に必要な語彙を缺き，十分な解析が行えない。そこで，こうした語彙を追加することにした。

3.3.3 その他利用可能な情報

月，干支，王朝名，元號，地名，人名等を追加した。

3.3.4 記號

記號をサポートするようにした。これにより，句讀點，括弧，下駄（〓）等の記號が混ざったテキストを認識する際の利便性が向上した。

3.3.5　除外リスト

取り込みたくないものを記述する除外リストをどう管理するかという問題がある。現狀では，除外リストは變換スクリプトに埋め込んだ形で管理しているが，複數人で作業することを考えれば，何らかのデータベース化が望ましいといえる。

3.4　品詞・素性

古典中國語と日本語は文法が全く異なるので，なんらかの品詞體系を用意する必要がある。3.2.1 節で述べたように，MeCab の辭書は第 5 欄以降に任意の素性を付けることができ，品詞に相當する情報はこれらの素性を使って記述するようになっているので，階層的な品詞體系を用いたり，意味素性を記述したりするようなことも可能である。

どのような品詞・素性を付けるかは，どのような處理をしたいかによる。MeCab は 1 つでも素性が異なれば別のエントリとして扱ってくれ[6]，同じ見出し語のエントリが複數あったとしても，コーパスを用いた學習等によって，適切なエントリを推定してくれるので，文脈から讀み取れるような情報であれば，素性として付ける價値があり得る。例えば，訓に關わる情報を付けておけば，訓を推定することが可能であり，これを使って自動訓讀システムを實現することができるかも知れない。あるいは，自動的に返り點を付けるシステムを作りたいのであれば，返り點を取るかどうかという情報を付ければ良いといえる。あるいは，日付，人名，地名等の情報を抽出するには，簡單な意味素性を付ければ良いかも知れない。

このようにいろいろな品詞・素性を付けることが考えられるが，多數の素性を付けるには辭書やコーパスを作成するための勞力が增える上，認識率が惡化するといえ，落しどころをどのへんにするかは問題である。おそらく，どうしても試行錯誤が必要となると思われるので，階層的な品詞・素性體系を用い，必要ならば素性を擴張するという方法が良いかも知れない。

著者は，幾つかの形式を試行錯誤した結果，

1．品詞 0（大品詞）
2．品詞 1（通常の品詞）
3．品詞 2（意味的素性 1）
4．品詞 3（意味的素性 2）
5．品詞 4（關係的素性）

6）素性のサブセットからなる內部エントリも作成するようである。

6．豫約

7．表記（正規形）

8．日本語表記

9．日本語での讀み

10．日本語での活用の種類

という形式を用いることにした。

3.4.1 大品詞

「品詞 0」素性（以下，『大品詞』と呼ぶことにする）は

n 名詞類

v 動詞類

p その他

からなる大雜把な品詞分類である。これは，もともと返り點付き漢文コーパスの利用を考慮して設けたものである。しかしながら，この情報は動賓構造に對應するようなものと看做すことができ，形態論的な情報だけでは決定しづらい通常の品詞（「品詞 1」素性）に比べて，コーパスや辭書の整備が不十分な状況において，頑健性を高める上で有效であると考えられる。

3.4.2 學習用コーパスからの變換

學習用コーパスの EOS のみの行を除く各行の形式は，3.2.1 節で述べた辭書の形式のうち，第 2, 3, 4 欄を削り，その部分をタブに置き換えたものになっている。よって，學習用コーパスの各行のタブを，0, 0, 0, に置き換えることで容易に辭書の形式に變換可能である。

3.5 動詞の品詞體系の問題

動詞の分類としては形態論・統語論的なものと意味論的なものが考えられるが，MeCab は隣接する形態素の素性の bigram を學習に用いているので，そこから導出できるような情報を動詞の品詞（素性）として付けても冗長であり，認識精度はあまり向上しないといえる。また，複雜な係り受け構造に關わるような情報は形態素解析のレイヤーでは扱えないので，統語論的な情報もまたあまり有益ではないと考えられる。古典中國語の動詞が屈折しないということも鑑みれば，意味論的な情報を中心に 2 階層の素性を用いて[7]動詞を分類するのが良いといえる。

7）MeCab の制約による。

動詞の分類としては，アスペクト（相），意志・無意志，視點などがあるが，古典中國語の動詞には時制やアスペクトによって變化せず（形態論的に表現されず），意志・無意志の兩方をとることがあり得，能動態・受動態の對立もない。特定のアスペクトや意志・無意志，視點をとるような動詞を見つけることができればそうした觀點での分類・素性付けを行う價値があるといえるが，一般にはあまり適切ではなさそうである。同樣の理由から槪念依存性に基づく分類もまたここでの目的には向いてないと考えられる。

一方，政治・司法・國家制度，軍事，雇用・役職，農業，商業・賣買・契約，家族關係，人間關係，文字・書物といった對象領域の分野による分類はある程度有用であると考えられる。しかしながら，この種の分野による分類を用いた場合，コーパスの分野依存性が高くなると考えられる。

現在のところ，アドホックに付けた品詞體系の方がある程度理論的・システマティックに付けたものより良い成績を收めており，適切な品詞體系を設計するためにはコーパスをもっと蓄積して，その係り受け關係や意味論的な分析を含めて檢討を行い，現實的な品詞體系の設計を行う必要があるといえる。

3.6　品詞體系のメンテナンス

コーパスに基づいて適切な品詞體系を設計できない段階においては，大まかな指針に基づきつつも，實際の文例に基づきコーパスを作る段階でアドホックに品詞・素性を付けるしかないといえる。この場合，問題となるのは品詞・素性の搖れである。例えば，同じ形態素に對して誤って違う品詞・素性を付けてしまうことがある。この問題の解決には，コーパスから生成した辭書が役立つと考えられる。もしコーパスから生成した辭書の同じ正規形に對して異なる品詞・素性を持つエントリーが存在する場合，品詞・素性の搖れが存在する可能性があるので，それをチェックすれば良い譯である。一方，異なる語彙閒の品詞・素性の搖れをチェックするには，辭書・コーパスで用いられている品詞・素性のリストが有用である。しかしながら，品詞・素性のリストだけで品詞體系を檢討すると個別の文脈の問題が捨象されてしまいがちであり問題がある。よって，品詞・素性のリストとその語彙の例，そして，その語彙が置かれたコーパス中の文を關連付けるようなツールが有用であると考えられる。

品詞體系を變更した場合，それに從って辭書やコーパスの品詞・素性を修正する必要がある。IPA 辭書から變換した辭書の場合，變換スクリプトを修正し，再度辭書を變換することで容易に品詞・素性を變換可能であるが，人手で作った辭書やコーパスに關しては工夫が必要である。この問題を輕減するためには，品詞體系を變更した時に辭書やコーパスを一貫して變更するようなツールを作るのが良いかも知れない。

また，別のアプローチとしては，形態素解析器の品詞體系に對して獨立した品詞體系，いわば情報の蓄積や交換のための標準的な品詞體系（ここでは，交換用品詞體系）を設け，コーパスや辭書はその品詞體系で書くというものである。形態素解析器の品詞體系に對しては交換用品詞體系から變換することにする譯である。交換用品詞體系が形態素解析器の品詞體系に對して十分に表現力があり，一意に變換可能であればこれは容易である。但し，實際には形態素解析に用いない品詞・素性までコーパスに書かなければならず，また，形態素解析の結果をそのままコーパスに用いることができないという點でこのアプローチには問題がある。

一方，MeCab の部分解析（制約付き解析）機能を利用するアプローチも考えられる。これはコーパスの內，變更され得る部分をワイルドカードにして，辭書の情報から補完するという方法である。この場合，元のコーパスの情報を落すだけであるから實現は簡單であるが，コーパスに記述していた情報の幾つかを落すことになるので，詳細な素性を付けたコーパスを蓄積している場合，問題である。

4　漢字の知識表現

漢字文獻の電子化を考える場合，その論理構造や內容に關わる側面と同時にその視覺的な構造に關わる側面の雙方を適切に扱うことが重要であるが，このためにはこの兩者の關係をどのように記述するかということが問題となってくる。いわゆる異體字の問題というのはそれを文字レベルで見たものだといえるが，このようなことは前述のように語彙レベルでも生じてくるし，文書の論理構造と媒體の視覺的・物理的構造の對應のようなレベルでもあり得る。ただ，その基礎となるのはやはり文字であり，文字をユニットとして文字によって構成されるさまざまな上位階層の問題を記述し得るといえる。

このように，論理構造や內容に關わる側面と同時にその視覺的な構造に關わる側面とのインターフェイスとして漢字を捉えた場合，視點や記述の荒さ / 細かさ等によって，幾つかのアスペクトを考えることができる。いわゆる『形』『音』『義』というのは漢字に對する視點の一種であるといえる。こうした視點で漢字を見た時，それぞれの視點での切斷面はしばしばグラデーションを作る。例えば，『形』の場合，字形（デザイン差）レベル，字體（文字の抽象的な視覺的表現）レベル，字體の包攝レベル，もっと包攝したレベル，……，を考えることができる。各レベルの境界がはっきりしていれば，各レベルを別のレイヤーとして分けて考えることができ，問題を單純化することができるのであるが，實際には，どこに境界を設けるかは一般には恣意的であるといえ，ある程度共通の規範がある部分もあれば，搖れている部分もあると考えられる。また，假に規範があったとしても，それは時

代や地域，コミュニティ等に依存し，變化するようなものと考えられる。結局，漢字には各視點（モーダル）毎に具象から抽象へのグラデーションを描く多次元空間上の場のようなものといえ，文字概念や觀念，書かれた文字，發音，觀念上の音，意圖，解釋，運用，といったさまざまな要素や現象等はその場の中の點や領域として捉えることができる。

漢字の知識表現とはこうした場を記述するということだといえる。この際，重要なのは，解釋や規範の搖れなどで變化しにくいものを基準に記述することと，互いに矛盾するような解釋や規範等であっても兩立するような枠組を用いることである。また，記述對象となる場は各アスペクトがしばしば連續的なグラデーションを描くようなものであり，あるいは，幸いにして明確な境界で切斷できるようなものだとしても各要素の組合せは無數に存在し得るようなものとなり得るが，記述というものは有限（でかつ，なるべく少數）のものでなければならない。このため，內包的記述と外延的記述をうまく組み合わせることが重要である。このことは言い替えれば，比較的客觀的に觀測される『書かれたもの』の有限個の記述と，それらを代表（標本點）する解釋や規範，抽象文字といった無限のインスタンスを內包する概念・觀念をどう關連付け，どう運用するかということである。著者が提案する Chaon モデルやそれに基づき著者らが開發している文字處理環境 CHISE ではこうした問題意識に基づき漢字を扱おうとしている。

4. 1　Chaon モデル

Chaon モデルでは文字は『素性』（feature）の集合として表現される。素性というのは『文字の性質』のことであり，素性の名前（素性名）と値の組で表現される。前述の文字のアスペクトは何らかの分節化と形式化を經て素性によって表現され得る。素性には，（1）文字間の關係を表現する『關係素性』，（2）文字の ID を表現する『ID 素性』，（3）その他，がある。文字をノード，關係素性をリンクとすることで文字間の關係のネットワーク（有向グラフ）が形成される。

文字は素性の集合で表現されると述べたが，素性の集合は標本點（インスタンス）としての文字の性質を記述したものと，素性の集合によって內包的に定義されその標本點を含むような文字の集合の 2 つの側面で解釋し得る。素性の集合で表現されるノード間の包含關係は素性の集合の集合演算で判斷できる（圖 2）が，それとは別に，關係素性を使って，さまざまなアスペクト，基準，規範，觀念等に基づく，

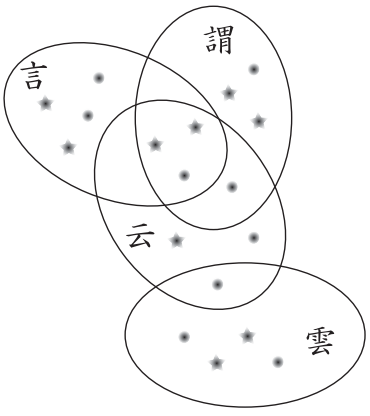

圖 2　素性の集合のベン圖

複數の包含關係を表現することもできる。

また，包含關係を表す關係素性を用いることで，素性の集合の差分的記述を行い，記述量を節約するとともに，可讀性を高めることも可能である。ここで，差分的記述のベースとなるノードのことを『親』といい，親から導出され，親に包含されるノードのことを『子』と呼ぶ。子は親が持つ素性を自分が陽に持っていない素性であっても自分が持つ素性と看做すことができ，これを『(素性の) 繼承』と呼ぶ。

4. 2 階層的素性名方式

汎用的な文字データベースを作る場合，用途や立場・學說などによって，文字素性の値に複數の選擇肢を設けたくなる場合がある。こういう時，單純に複數の値が記述できるだけでなく，各々の値の出典情報などのメタデータも付加したいことが少なくない。こうした場合，文字素性の値か名前のどちらかを構造化する必要がある。『階層的素性名方式』というのは後者の方法の一種である。

階層的素性名方式は構造化の對象となる文字素性の名前（文字素性基底名）に値を選擇するための識別子（『ドメイン識別子』と呼ぶ）を付けた文字列を生成し，それを名前（文字素性具象名）として用いたり，同樣にメタデータ識別子を付けた文字列を生成しそれを名前（文字素性メタデータ名）として用いる方法である。

この名前は次のような規則で生成される：

文字素性具象名
：＝文字素性基底名@ドメイン識別子
文字素性メタデータ名
：＝文字素性具象名＊メタデータ識別子
｜ 文字素性メタデータ名＊メタデータ識別子
｜ 文字素性メタデータ名@ドメイン識別子
ドメイン識別子
：＝基底ドメイン識別子
｜ ドメイン識別子 / 基底ドメイン識別子

例えば，總畫數を表す文字素性名を total-strokes とし，ドメイン識別子として ucs を用いる時，文字素性具象名は total-strokes@ucs となる。また，出典情報を表すメタデータ識別子を sources とする時，total-strokes@ucs の出典情報は

total-strokes@ucs＊sources

で表される。

部首と部首內畫數のように異なる種類の文字素性の値が對應關係を持っている場合，ド

メイン識別子を用いてその對應關係を表すことができる。例えば，部首を ideographic–radical，部首內畫數を ideographic–strokes で表す時，

ideographic–radical@ucs

ideographic–strokes@ucs

の兩者は對應する。

値を構造化する手法と名前を構造化する手法を比べた場合，前者はドメイン識別子を必要としないという利點を持っているものの，値が構造データとなるので C のような單純な記憶管理機構しかない環境では不便である。また，高速化を要するような單純な處理の場合，大抵，複數の値やメタデータを必要としないといえる。また，Chaon モデル的には文字が文字素性の單純な集合になっている方が自然であり便利であるが，値を構造化すると複數の値同士の集合演算を要し，處理が複雜になる。このようなことを鑑み，現在の CHISE では共有文字データベース內では原則として値ではなく名前を構造化する方針を採っている。

4. 3　文字オントロジーの構成法

Chaon モデルはメタなモデルであり，これに基づく具體的な文字データベースの構成法や文字處理システムの實現の仕方にはさまざまな形がありえる。このことは文字に關するさまざまな觀念や概念を記述可能であることを意味しているといえるが，とはいえ，現實に文字を處理するためには，具體的なアーキテクチャや文字知識を記述するための指針が必要だといえる。

そこで，現在の所，CHISE project では『(社會的に) 共有される文字觀念』に相當する文字知識を記述した文字オントロジー (共有文字データベース) を中心にした文字處理アーキテクチャを採用している。この文字オントロジーとして「CHISE文字オントロジー」の編纂を續けている。

CHISE文字オントロジーはさまざまな文字觀念を差分的に記述するためのベースになるような汎用的な文字オントロジーを提供することを目指している。このため，字形・字體の細かな差異を捨象した抽象的文字觀念，Unicode や JIS X 0208 のような各種符號化文字集合における符號化文字，さまざまな文字符號の規格や辭書などの文字表における例示字體・字形の情報など，抽象・具象のさまざまなレベルの代表的文字概念を收錄している。

これらの各レベルは主に字體差 (比較的大きな形の差異) を示す ⇀ *denotational*，↽ *denotational* 素性と比較的小さな字體差・字形差 (比較的小さな形の差異) を示す ⇀ *subsumptive*，↽ *subsumptive* 素性を用いて，文字オブジェクト間の繼承關係として記述してい

る。卽ち,

抽象的 ⇒*denotational* 具象的

や

抽象的 ⇒*subsumptive* 具象的

という風に文字間の繼承關係を記述する譯である。⇒*denotational* と⇒*subsumptive* は混在して使うことができ,

大粒度抽象文字⇒*denotational* 中粒度抽象文字⇒*denotational* 細粒度抽象文字

⇒*subsumptive* 字體⇒*subsumptive* 抽象字形

などのように多段的繼承關係を記述することも可能である。

漢字の場合,抽象・具象關係は形・音・義のそれぞれに存在し得るといえるが,今の所,基本的に形の側面に基づくものしか扱われていない。しかしながら,他の要素に基づく抽象・具象關係も將來的には記述したいと考えている。形・音・義それぞれの抽象・具象關係が一致しない場合でも,4. 2 節で述べた階層的素性名方式に基づき,それぞれを別ドメインとすることで多元的な繼承關係を記述することができる。この場合,どのドメインの繼承關係を用いるか(あるいは用いないか),優先させるかといったことはアプリケーションとドメインを對應づけることによって制御可能である。

5　漢字處理の多層化

5. 1　文字と形態素

古典中國語においては 1 文字からなる形態素が少なくなく,文字と形態素はさまざまな面で形式的に重なって見えるといえる。形態素辭書における各種素性は,形態素が 1 文字の場合,文字素性の一種と看做すことが可能であり,實際,字義や發音に關わるような情報は漢字辭書に記載される主要な項目のひとつであるといえる。品詞をはじめとする文法的な情報もまた同樣に考えることができるだろう。

漢字處理において重要な問題のひとつである異體字の問題もまた文字と形態素の雙方の領域にまたがる問題のひとつといえる。例えば,常用漢字を中心とする現代日本語表記における漢字(以下,『新字』とする)をいわゆる『康熙體』を中心とする傳統的な漢字(以下,『舊字』とする)に變換する場合,文字單位に新字を舊字に變換するのでは一意に決定できなくなったり不適切な變換をすることになる譯だが,このことは異體字の問題の幾つかは形態素や語彙の世界における表記の搖れとして捉えなければならないことを意味している。漢字の異體字關係は時代や地域,分野,テキスト,文脈等に依存するということが知られているが,こうした現象をきちんと捉えるためには,文字の世界だけに閉じて記述

することはできず，3.2.2 節で述べた形態素解析のための文法コーパスのように，對象となる文字 / 形態素をそれが出現する文脈を含んだ形で記述する必要があるといえる。形態素解析のための辭書やコーパスと文字オントロジーを連携させるということは，形態素解析にとっても有用なことであるが，文字處理の側から見ても重要なことだといえる。

2 文字以上からなる形態素の場合，それを文字と看做すことはできない譯であるが，素性の集合で文字を表すという Chaon モデルの方法は，實のところ，概念一般に對する知識表現の一種に他ならなく，文字に限定されるものではないので，文字の場合と同様のやり方で形態素のオントロジーを記述することは可能である[8)]。もし，CHISE の文字オントロジーと同様な方法で形態素のオントロジーを構成すれば，文字と形態素の差異を考慮しつつ，兩者をシームレスに扱うことが可能になると考えられる。

5.2　グリフ・コーパス

グリフに關わるような問題，例えば，どういう時にどういうグリフの差異が別字として書き分けられ辨別されるのかだとか，どういう時に對應する異體字と看做されるのか，あるいは，グリフの規範意識といったものを考える場合，そのグリフが用いられた文脈を考慮する必要があるといえる。このためには，グリフのためのコーパス（これを『グリフ・コーパス』と呼ぶことにする）が有用である。

グリフ・コーパスというのはグリフを用いられたテキストと對應付けた形でデータ化したものであり，透明文字付き文字畫像［5］や畫像マークアップされた文字畫像［6］というのはその一種と考えることができる。文字列としてのコーパスと對比させて考えれば，前者はプレイン・テキストに相當し，後者はマークアップ・テキストに相當すると考えられる。單なる文字畫像もグリフ・コーパスの一種と考えるが，現在の古典中國語テキストに對する文字認識技術を前提にした場合，文字に分節化されていないままでは精度の點で難があるといえ，文字單位に切り出され，その切り出された各文字の位置關係が判り，また，各文字がどういうグリフであるかということを示し得るような情報が付與されていることが望ましい。こういう觀點からいえば，檢索のためになるべく異體字を正規化した透明文字付き文字畫像はグリフを指し示す際の精度という點で難がある。（異體字を正規化した）抽象文字はグリフにとって重要な素性のひとつといえるが，同じ抽象文字に對應するグリフ間の差異を指し示すことができない。こうしたものは現狀では基本的に目で

8）MeCab の辭書やコーパスの形式も（制限があるとはいえ）素性の集合で表現されたものである。

見て判斷するしかなく，大量のデータに對して機械的に處理することが困難である。

グリフを表現するための素性としては，文字認識で用いられるヒストグラムといった特徴量もあるが，この種のものは機械可讀ではあるものの，人間にとっての可讀性が低く，したがって，人間にとっての意味を必ずしも反映しないといえ，グリフを辨別し指示するための素性としてはあまり適切ではないと考えられる。

一方，IDS（Ideographic Description Sequence）［1］のような漢字の部品の組合せ方を表現したもの（これを『漢字構造表現』と呼ぶことにする）がある。これはグリフ表現という觀點では必ずしも全ての漢字を表現できる譯ではないという點で問題があるものの，多くの漢字を表現することができ，また，人間にとっての理解に近いという點で優れている。本來の IDS は部品として UCS の統合漢字および漢字部品を用いることになっているが，漢字構造表現自體は必ずしもそれに限定されるものではなく，部品を指し示すことができるならばどのような部品を使っても成り立ち得るといえ，實際，CHISE では CHISE の文字オントロジーにある任意の文字オブジェクトを利用できるように擴張している。［7］ この擴張された漢字構造情報では，部品として抽象文字をとることもできるし，字形レベルのものをとることができ，異なるレベルの部品を混在することもできる。この結果，グリフ間の包攝關係やどの部分の微細な差異に着目しているのかといったことなども表現できる。

また， Adobe Japan 1 などのグリフ集合や異體字が細かく區別された符號化文字集合を利用してグリフを指定することも考えられる。この場合，グリフを細かく區別しようと思えば思う程，表現できないグリフが增えることになるので，スケーラビリティの點で問題がある。しかしながら，多數流通する PDF文書等をグリフ・コーパスとして利用する際には，Adobe Japan 1 をはじめとする CID は有用な情報のひとつであると考えられる。

5. 3　グリフ・コーパスの形態素解析

5. 1 節や前節で述べたような文字處理と形態素解析の連携やそのためのオントロジーの統合といったことは，グリフと形態素に對しても成り立つといえる。グリフに關わる規範意識やグリフ間の對應關係，どの差異に着目するかという問題はしばしば單一の文字だけで論じられるものではなく，グリフが用いられた文脈や文法的，意味的，視覺的構造なども含めて考える必要があるといえる。こうしたことを鑑みれば，文字レベルだけで考えるのではなく，形態素をはじめとする上位層の構造を含めた表現や處理が有用なのではないかと考えられる。そのための第一步として，グリフ・コーパスの形態素解析について考えてみる。

グリフ・コーパスの形態素解析は，どこまで細かくグリフを區別するかという問題を無

視すれば，透明文字付き畫像を形態素レベルの文法的な情報を含む文字畫像マークアップ・テキストに變換する問題という風にとらえることができる。これは基本的にはグリフの位置情報を管理しながら形態素解析を行うことで實現できるといえる。

前節で述べたように，グリフを表現する手段は複數考えられるが，これは文字オントロジーの場合と同樣に，それぞれを素性として表現し，その組合せでグリフを表現することによって，これらを併用したり，ある情報から別の情報を取り出したりといったことが可能になるといえる。こうした方法によってグリフの知識表現を記述した『グリフ・オントロジー』を構成することができるが，これは實のところ，文字オントロジーの一部をなすものといえる。

5. 1 節で議論したように，文字レベルのグリフと同樣に，形態素レベルのグリフ（列）の知識表現を考えることができるが，これがグリフ・コーパスの形態素解析器における辭書に相當するものといえる。そして，文法的文字畫像マークアップ・テキストが文法コーパスに相當するものと考えられる（これを『文法グリフ・コーパス』と呼ぶことにする）。

結局，グリフ・オントロジーと文法グリフ・コーパスを蓄積することによって，グリフ・コーパスの形態素解析が實現できると考えられる。當然のことながら，この情報はこれらがその一部として含むグリフの文字レベルでの頻度情報や文脈的情報，通常の形態素解析のためのデータ，それに關わる異體字情報等を含むといえる。

6　お　わ　り　に

常用漢字改訂に關わる議論でも明らかなように，漢字を考える上で字體・字形（ここではこの 2 つを總稱して『グリフ』と呼ぶことにする）の規範に關わる問題は避けては通れないといえる。しかしながら，グリフおよびその規範意識に關わるようなさまざまな要素を機械可讀に表現したデータベースやオントロジーの類はこれまでほとんど作られてこなかったといえる。もちろん，字形情報自體は畫像データとして盛んに電子化されてきているし，そうしたデータベースは多いとはいえないものの存在していて，實證的文字研究をする上での重要なツールとなっている。しかしながら，視覺的な情報としての字形情報に關するメタデータ，例えば，どこがどう違っているのか？　どういう意識で變えたのか，どれが似ていてどれが似ていないのか（違和感を感じるのか），といった視覺的な情報の構造やその意味に關わるような情報の機械可讀化はあまり進んでいないといえる。そのため，こうした事項は人間が目で見て判斷するしかないが，漢字は文字數が多くこうした作業は大變な手間である。

また，グリフが置かれた文脈情報まで含めて扱うことは難しく，こうしたことから，1

文字単位の頻度情報に頼ることが多くなってしまったといえる。しかしながら，書記言語もまた自然言語の一種であると考えられ，書記系としての構造（文法）を持ち，出現頻度の少ないものが重要な意味を擔い得るような性格を持っていると考えられる。こうしたことから，機械可讀な（構造化された）グリフ・コーパスを實現し，書記言語としての構造に卽したグリフ情報の分析を行うことがグリフに關わる問題を考える上で重要だといえる。

本論文では，このような問題意識に立ち，著者がこれまで取り組んできた文字やグリフの知識表現や古典中國語のための形態素解析技術について概說するとともに，これらを統合した多面的な漢字知識處理を提案した。これはいわば漢字の『形』と『義』の兩面を同時に扱うための方法論のひとつであるといえる。ここでは文字と形態素という 2 つの階層のみを扱っているが，古典中國語のための構文解析や意味解析を實現することで，より上位の階層についても扱っていくことが考えられる。また，ここでは『音』について扱っていないが，これもまた文字オントロジーの記述やテキスト解析において重要な要素のひとつであるといえ，今後取り組むべき課題のひとつである。

古典中國語テキスト處理をより高度な形に進展させるためには，入力したデータが新たなデータを生み出すような，データの整備が生產性を高めるような體制が望ましいといえる。文字や形態素のオントロジーや文法コーパス，グリフ・コーパスといったものは，『データを生み出すことができるデータ』の一種といえ，データの整備の進展とともにより『賢い』機械處理を實現していくことができると考えられる。こうしたことを鑑みれば，こうした基盤データの整備を行うことにより機械處理のツールチェーンの足りない部分を埋めていくことが，これからの人文系情報處理にとって重要な課題だと思われる。

參 考 文 獻

［1］ International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set* (*UCS*), 2003 年 3 月. ISO/IEC 10646 : 2003.

［2］ MeCab : Yet Another Part-of-Speech and Morphological Analyzer. http : //mecab.sourceforge.net/.

［3］ MORIOKA Tomohiko. CHISE : Character Processing Based on Character Ontology. In Takenobu Tokunaga and Antonio Ortega, editors, *Large-Scale Knowledge Resources*, Vol. 4938 of *LNAI*, pp. 148－162. Springer, 2008 年.

［4］ 永崎研宣. 人文科學のためのデジタル・アーカイブにおけるステイクホルダー —— 佛教文獻デジタル・アーカイブを手掛かりとして ——. 人文科學とコンピュータシンポジウム論文集, pp. 347－354. 情報處理學會, 2007 年 12 月.

［5］ 安岡孝一. 透明テキスト付き畫像へのいざない. 東洋學へのコンピューター利用第 14 回研究セミナー, 京都大學學術情報メディアセンター第 71 回研究セミナー, pp. 31－42, 2003 年 3 月.

［6］ 守岡知彦．文字畫像のマークアップの試み．東洋學へのコンピューター利用第 14 回研究セミナー，京都大學學術情報メディアセンター第 71 回研究セミナー，pp. 21－30，2003 年 3 月．

［7］ 守岡知彦．CHISE 漢字構造情報データベース．東洋學へのコンピューター利用第 17 回研究セミナー，全國文獻・情報センター人文社會科學學術セミナーシリーズ，京都大學學術情報メディアセンター第 78 回研究セミナー，pp. 93－103，2006 年 3 月．

［8］ 守岡知彦．Concord：プロトタイプ方式のオブジェクト指向データベースの試み．Linux Conference 抄錄集 Vol. 4，2006 年．

［9］ 守岡知彦．文字オントロジーに基づく文字處理について．情報處理學會研究報告 Vol. 2006，No. 112，pp. 25－32，2006 年 10 月．2006－CH－72．

［10］ 守岡知彦．MeCab を用いた古典中國語の形態素解析の試み．情報處理學會研究報告 Vol. 2008，No. 73，pp. 17－22，2008 年 7 月．2008－CH－79．

［11］ 中田充，寶珍輝尙，都司達夫．サイエンティフィックデータベースのためのデータモデルの一提案 ～考古學データベースを例として～．情報處理學會研究報告 Vol. 1995，No. 12，pp. 65－72，1995 年 1 月．1995－DBS－101．

［12］ 中田充，寶珍輝尙，都司達夫．名前付き集合モデルを用いた DREAM モデルの定義．情報處理學會研究報告 Vol. 1997，No. 38，pp. 1－8，1997 年 5 月．1997－DBS－112．

［13］ 白須裕之．人文系データベースを構築するとはどういうことか？ 漢字文獻情報處理研究 Vol. 9，pp. 11－19，2008 年 10 月．