古典中国語形態素コーパスの Linked Data 化の試み

守岡 知彦
京都大学人文科学研究所

古典中国語のための文字としての漢字は 1 字が概ね形態素に相当する表語文字の一種であり、各字は基本的に形態素を表現している。そのため、例えば、異体字処理の場合、文字層だけでは処理できず、形態素や語彙に関する知識が必要となる場合がある。よって、漢字の知識表現も一般には文字層だけで閉じた記述を行うことには問題があるといえ、形態素層との連携した複合的な記述を行うことが望ましいといえる。そこで、現在開発中の MeCab 用の古典中国語形態素コーパスを Linked Data 化し、CHISE 文字オントロジーと統合することを試みた。

Linked Data for Morphological tagged Classical Chinese Corpus

MORIOKA, Tomohiko
Institute for Research in Humanities
Kyoto University

Chinese characters, as the script of the classical Chinese language, are logograms, namely each abstract character basically indicates a morpheme. Therefore some character processings, such as variant processing, can not complete only character layer, but also require higher layers such as morpheme and word. Knowledge representation of Chinese characters also should be integrated with knowledge representation of classical Chinese morphemes. To satisfy this requirement, I tried to convert a morphological tagged classical Chinese corpus, now we are developing as a part of morpheme analyzer for classical Chinese based on MeCab, to Linked Data, and integrated with the CHISE character ontology.

1 はじめに

古典中国語のための文字としての漢字は 1 字が概ね形態素に相当する表語文字の一種であるといわれる。このため、漢字辞書での記述や文字処理のためのデータ等での漢字に関する情報において、文字としての情報と形態素としての情報が区別されず渾然一体となっていることも少なくない。例えば、異体字・類字関係には文字単位のものと形態素単位のものがあるが、後者は文字単位の処理だけでは正しく扱うことができない（例：自動変換できないことがある）。また、文字の使用実態を調べる場合も、単純に文字単位の頻度を見るだけでは不十分なことがあり、どういう形態素を表現するためにどういう文字が使われているかを調べることは重要であるといえる。

一方、我々は古典中国語のテキスト解析のために MeCab[4] 用の形態素コーパス [9] を開発しているが、この開発中の形態素コーパスを効率的に管理するためにはコーパスと辞書と形態素レベルでの異体字関係を効率的に管理するためのツールが必要であるといえる。古典中国語の形態素は、日本語と違って、屈折がなく、品詞の転用が多く、多く形態素が 1 文字であるという性質があり、一般的な品詞だけでなく、意味的なカテゴリーを含んだ品詞体系（語彙カテゴリー）を設けることが望ましいが、どういうカテゴリーを設けるのが良いかは自明ではなく、入力のし易さと体系性や言語学的な適切さ等を

バランスさせた品詞体系を開発するためには、初期段階では揺れを抑制するよりも、複数の入力者の知識や言語感覚を利用し、各入力者の入力結果を比較・分析するという方法を取らざるを得ないが、このためには異なる品詞体系の辞書や形態素コーパスが共存可能であることが望ましい。

こうした問題を解決するため、MeCab 用古典中国語形態素コーパスの Concord[6] / EgT[7] を用いた Linked Data 化を試みた。また、この形態素 Linked Data と CHISE 文字オントロジー [5] を互いにリンクさせ、文字と形態素の 2 層からなる漢字オントロジーの実現を試みた。

2 関連研究

2.1 WordNet

WordNet は英語の語彙を synset と呼ばれる同義語のグループに分類し、synset 間の関係を記述した機械可読な概念辞書である。WordNet は BSD ライセンスによって公開されており、自由に利用可能であり、さまざまな派生物も作られている。

英語以外の言語に対する WordNet を開発する試みも多数行われており、中国語においても幾つかが開発されたようである。しかしながら、中国語を扱う場合、漢字に関わる幾つかの問題が生じる。そのひとつは外字や符号化に関わる問題であるが、より本質的な問題としては、書記言語としての中国語を考える場合、同義語や類語・語彙間の関係と漢字の異体字・類字関係の双方を考える必要があり、それらの全体の系は字形・字体・文字・形態素・語彙といった領域にまたがったものとなってしまうからといえる。これに対し、単語という視点だけで記述しようとする WordNet は文字という視点だけで記述しようとする従来型の多くの文字データベースと同様に適切な記述を行う上で問題があるといえる。

2.2 Hantology

Hantology[1][2] は WordNet の問題を解決するために開発された中国語における漢字のオントロジーである。Hantology は漢字の形音義に関する情報、CDP 方式による部品の組み合わせ方に関する情報（漢字構造情報）や説文解字の情報、上古音、中古音、現代音といった発音の情報、字義、品詞、異体字情報、その字を含む単語の情報等を文字を単位に記述しており、上位オントロジーとして SUMO (Suggested Upper Merged Ontology) を用い、OWL によって符号化している。約 5 万文字を収録している。

Hantology は説文解字や広韻、漢語大字典といった辞書の規範的な記述をベースにしているようであり、規範的な用例は含むものの、用例ベースとはいえず、コーパスと紐付けられている訳ではない。どちらかといえば、典型的な漢字辞書の記述を機械可読化したようなものと見ることができ、結果的に、現代中国語と古典中国語の情報が十分に区別されないまま混在したものとなっている。

2.3 CHISE 文字オントロジー

CHISE 文字オントロジー [5] は文字処理のために著者らが開発している軽量オントロジーである。CHISE 文字オントロジーは Unicode に収録された文字の情報の他に、漢字に関しては Unicode の包摂規準以外に超抽象文字や字体・字形といった複数の包摂粒度による漢字のグリフに関わる情報を持っている。各漢字には、部首・画数や異体字・類字関係等の情報、IDS 形式 [3] に基づく漢字構造情報、各種文字符号でのコードポイントの情報、各情報の出典等のメタ情報を収録しており、現在のデータ総数は約 24 万オブジェクト（抽象文字、超抽象文字、字体、字形等の各粒度のオブジェクトののべ数)、89 万トリプルである。

CHISE 文字オントロジーは、現在の所、文字に関わる情報だけを収録しており、文字以外のリソースは単なる識別子や外部へのリンクになっている。

2.4 ChaKi

ChaKi（茶器）はタグ付きコーパスを管理・検索するためのツールである。MeCab やCaboCha 等で処理したタグ付きコーパスを関係データベース管理システム (RDBMS) で管理し、さまざまな検索機能や修正機能を提供している。

ChaKi の旧版は RDBMS として MySQL を用いており、ネットワーク上でコーパスを共有することが可能であったと思われるが、新版である ChaKi.NET では SQLite を用いておりスタンドアロン動作のみとなっている。また、両者とも Windows でしか動作しない。

形態素コーパスの入力支援システムとして ChaKi を見た場合、Unicode をサポートしており、現代日本語以外の言語も利用可能であるといえるが、外字の管理には別途工夫が必要であるといえる。形態素の属性は現代日本語用を想定した基本 9 種に固定されており、これらにマップできない情報を表現するためには custom フィールドを用いる必要がある。但し、この custom フィールドの情報は検索対象とはならないようである。また、ChaKi は係り受け情報付きコーパスをサポートしている。また、形態素コーパスに関しては複数の文の一括修正が可能である。

ChaKi は異なる品詞体系をサポートしたり、品詞体系間の関係を管理する機能はないようなので、あらかじめ品詞体系やガイドラインを定義した上で、用例を見ながらコーパスを修正するような使い方が想定されているようである。

また、マルチユーザー機能や版管理機能等はサポートしていないため、複数人からなるグループでコーパス開発を行う場合、別途なんらかのツールが必要になると考えられる。

3 目標

3.1 漢字の知識表現としての側面

漢字の性質を記述する上で、各時代の規範的な辞書の情報を機械可読化することにも一定の意味はあるが、この種の規範的な情報はある種の思想史的な情報を含んでしまい、漢字の性質自体や実際に漢字の運用実態に関する情報としてはやや問題があるといえる。こうしたことを考えれば、用例となるコーパスと紐付けることが重要であるといえる。

漢字の文字としての側面と形態素としての側面という二面性を考えた場合、漢字の性質に関する情報は文字か形態素のどちらかではなく、その両面から記述することが望ましいといえる。コーパスも同様に両者の側面を適切に表現し、この両面から漢字の知識表現とリンクすることが望ましいと考えられる。

また、用例ベースが望ましいということは、(運用面はさておき原理的には) 用例がいくらでも追加でき、場合によっては互いに矛盾する情報を適切に混在できることが望ましいということであり、複数の包摂規準や品詞体系・形態素の認定基準等が共存できるような枠組が望ましいといえる。

3.2 形態素コーパスの管理機能

古典中国語用の形態素解析器で必要となる形態素辞書と形態素コーパスの開発において、その入力作業における形態素認定基準や品詞体系等の『揺れ』の問題に対処するためになんらかのツールを用いてそのリファクタリングを支援することは重要である。

こうした揺れの問題に対処するための方法としては、ChaKi が行っているように、判断に迷った時に他の用例を提示するなどして、入力時に揺れが起こりにくくするというアプローチが考えられるが、このためには品詞体系や形態素の認定基準等をあらかじめきちんと決めておくことが実用上不可欠と考えられる。つまり、我々が行っている古典中国語用の形態素コーパスの開発の場合のように、きちんとした品詞体系がなく、絶えず試行錯誤することとなり、結果的に、プロジェクトの途中で形容詞を廃止するというようなドラスティックな変更を行うようになった場合、入力時に揺れを小さくする工夫は単に入力効率を下げるだけの結果になりかねず、むしろ、揺れを含んだコーパスをなるべ

く少ない労力で整理するための後処理用のシステムが有用であると考えられる。

大量のコーパスを処理することを考えた場合、自明なケースに関してはなるべく自動的に対応関係を推定し、人間の判断を必要とする部分を小さくするような工夫を行うことが望ましい。また、古典中国語のような古典語のテキストを対象とする場合、言語学的な専門知識を要する部分と内容に関する専門知識を要する部分と高度な専門知識を要さずに判断できる部分が混在するので、入力者だけでは判断が難しい高度な専門知識を要する部分を切り分けられるような仕組みが望ましいといえる。また、人間が行った判断を、該当個所だけでなく、同様な例に対して自動的に適用できる仕組みがあることも望ましいといえる。そのためには、単なる一括置換機能では不十分であり、修正過程を一般化・知識表現化し、（プログラムとして）再利用可能となるような仕組みがあることも望ましいといえる。そして、こうしたメタな修正過程に対する修正が可能であることも望ましい。

4　設計

漢字の表語文字としての側面を鑑みれば、文字と形態素という２種類のオブジェクトを設定し、文字オブジェクトと形態素オブジェクトの間に『指示する／指示される』関係のリンクを張るのが良いと考えられる。これは漢字を構成する３要素といわれる『形音義』の内、『形』を中心とした情報を文字オブジェクト、『音義』を中心とした情報を形態素オブジェクトで表現するものという風に見ることができる。

即ち、文字オブジェクトは視覚的形状に関して、その抽象・具象関係や異なる字形間の同値性に関わる情報や文脈依存性、差異の条件、漢字構造情報、出典情報や用例、例示字形や文字・グリフ符号の情報、文字単位での異体字・類字情報等を記述したものであることが望ましいといえる。一方、形態素オブジェクトはその形態素の正規形（各種規範における『正字』等）と各種表現形（実際にテキストに出てくる形；異体字・類字等）、品詞情報、発音、その他形態論的情報等で表現できる。

ここで、形態素オブジェクトの正規形や表現形は文字オブジェクト、もしくは、その列として表現できる。１文字の時と文字列の時での場合分けを避けて簡単にするために、これらは常に文字オブジェクトの列によって表現することにし、そのために『見出しオブジェクト』を設けることにする。つまり、形態素の正規形や表現形の情報は符号化文字列によって表現するのではなく、見出しオブジェクトによって表現する訳である。これにより、Unicode に存在しない文字であっても CHISE の手法を用いて表現することができ、また、正規形や表現形を適切な包摂粒度の文字（グリフ）を用いて記述することができる。また、ある見出しオブジェクトを正規形、もしくは、表現形として持つ形態素オブジェクトの集合を簡単に抽出できる。

一方、複数の品詞体系をサポートすることを鑑みた場合、品詞もまたオブジェクト化することが望ましいといえる。我々が開発している古典中国語形態素コーパスでは大品詞、品詞と２階層の意味カテゴリーからなる４階層[1]の素性で品詞（語彙カテゴリー）を表現するようになっているが、この４素性をひとまとめにしたものをオブジェクト化すれば良いといえる。これにより、同様な語彙カテゴリーを示すと考えられる品詞オブジェクト間にそのことを示す関係リンクを張ることで、漢字の異体字処理と同様な方法で異なる品詞の同値性を示すことができる。また、カテゴリーの包摂範囲の差も is-a 関係や『包摂する／される』関係で表現できる。品詞体系は品詞オブジェクトの集合で示すことができる。具体的には、品詞体系を示すラベルを定義し、それを形態素オブジェクトの素性の１つとして表現すれば良い。

文は、文字単位で見た場合には文字オブジェクトの列と看做すことができ、また、形態素単位で見た場合には形態素オブジェクトの列と看做すことができる。この二面性を実現するために、文は『文オブジェクト』と見出しオブジェクトと形態素オブジェクトの列の三者の関係によって表現することにする。

[1]初期には５階層のものも用いていた。

ここで、文オブジェクトはあるテキストのどこかの場所にある文を示すものとし、抽象的な文ではなく、実際に書かれた文としての『文の出現』（インスタンス）を示すものとする。よって、その ID はテキストの ID とテキスト中の位置情報によって生成することが望ましいといえる。

これに対し、見出しオブジェクトと形態素オブジェクトの列は抽象的な文を示すものとなる。よって、文オブジェクトから見出しオブジェクトや形態素オブジェクトへのリンクは書かれた文の解釈や校訂といったものに対応するものといえる。これにより、例えば、ある書かれた文に対して異なる解釈が存在する時に、ある文オブジェクトから複数の見出しオブジェクトや形態素オブジェクト列へのリンクを張ることでこうした揺れを表現することができる。

見出しオブジェクトは形態素のものと文のものとで別にすることも考えられるが、ここでは同じ文字列を示すものは同一のオブジェクトにまとめる方針を採用する。

短単位と長単位のような、複合語や固有名詞等での形態素の認定基準の揺れを表現するために、『句オブジェクト』を導入する。これは文オブジェクト同様に、見出しオブジェクトと形態素オブジェクトの列へのリンクを持つオブジェクトとするが、文オブジェクトの場合と異なり、あるテキスト中でどこかの箇所で実際に書かれた『句の出現』を示すのではなく、抽象的な句を示すものとする。

形態素コーパスの情報を修正する場合、形態素オブジェクト（あるいは、句オブジェクトや見出しオブジェクト等）自体を書き換えることも可能ではあるが、人間が行った修正過程からより一般的な修正法を推論して実行したり、あるいは、この推論の結果が望ましくない場合に制約条件を付けるなどして推論プログラムを修正する『メタな修正過程』を実現することを考えた場合、むしろ、誤りを含んだものや旧形式のものも含めて、原則として、あらゆる形態素オブジェクトはそのまま保持し、その代わりに、書き換えを示すリンク情報だけを付与することにする。Linked Data として考えた場合、これは誤ったものや過去のものも含め、形態素オブジェクトや形態素コーパス中の各文に対しユニークな URL を与え、Web 上から参照できることを意味する。これにより、編集途中の任意のスナップショットにリンクを張ったり、複数のスナップショット間の関係や遷移過程を、人間・機械を問わず、第三者もデータ化することが可能となる訳である。

5 実装

手間を省くために、文字オブジェクトとして CHISE 文字オントロジーをそのまま利用した。また、CHISE 文字オントロジーとの統合を容易にするために、形態素オブジェクトや品詞オブジェクト、見出しオブジェクトの格納にも CHISE 文字オントロジーの基盤となっているグラフ型データベースである Concord[6] を用いた。

そして、XEmacs CHISE 上の Concord の Emacs Lisp バインディングを利用して、MeCab 用の形態素コーパス・ファイルや辞書ファイルを Concord 内に取り込むツールや Concord 内に格納された形態素 Linked Data を MeCab 形式のコーパス・ファイルとして出力するツール等を開発した。また、形態素 Linked Data の情報を利用して揺れを推定したり自動的に修正するためのツールも開発している。

5.1 データ構造

Concord では各オブジェクトは “genre” と呼ばれるオブジェクトのタイプのようなものを持つことになっており、古典中国語形態素 Linked Data では表 1 に示す genre を用いている。

種類	Concord での genre
文字	`character`
見出し	`entry@zh-classical`
形態素	`morpheme@zh-classical`
句	`phrase@zh-classical`
文	`sentence@zh-classical`
品詞	`word-class@zh-classical`

表 1: オブジェクトの Concord genre

各オブジェクトは、素性名と値の対（素性対）の集合によって表現される。素性対は RDF におけるトリプルに相当するもので、同様に、素性名は述語、素性値は目的語に相当するものである。

素性名は RDF での述語に変換することができる。この詳細に関しては [8] で述べているので、ここでは省略する。

異なる品詞体系や認定基準や解釈等を扱うために、必要に応じて、階層的素性名を用いる。また、古典中国語を表すためのドメインとしては `zh-classical` を用いることにする。

5.1.1 文字オブジェクト

文字オブジェクトは CHISE 文字オントロジーのものをそのまま用いることにしている。

但し、文字と形態素の関係を表現するために、文字オブジェクトに素性

`*instance@morpheme-entry/zh-classical`

を付与し、ここに形態素の見出し文字列オブジェクトの集合を格納している。

5.1.2 見出しオブジェクト

表 2 に見出しオブジェクトの素性一覧を示す。

素性名	値の内容
`=id`	ID
`=name`	名前（対応する文字列）
`character`	対応する文字へのリンク
`<-entry@morpheme`	この文字列を表現形として持つ形態素の集合
`<-entry@morpheme /canonical`	この文字列を正規形として持つ形態素の集合
`<-entry@phrase`	この文字列を表現形として持つ句の集合
`<-entry@sentence`	この文字列を表現形として持つ文の集合

表 2: 見出しオブジェクトの素性

`=name` はその見出しオブジェクトが表現する文字列を示す。`=id` はこれを変換して作ることにしている。

見出しオブジェクトが表現する文字列が 1 文字の場合、対応する文字オブジェクトを素性 `character` で表現する。但し、2 文字以上の場合は省略する。[2]

`<-entry@morpheme`(*/canonical*) は形態素オブジェクトから見出しオブジェクトに張られたリンクによって生成される逆関係素性である。サブドメイン `canonical` が指定されたものは正規形を示し、省略されたものは表現形を示し、両者ともそれぞれの関係に従い、その見出しオブジェクトを表現形もしくは正規形として持つ形態素オブジェクトの集合を値として持つ。但し、表現形と正規形が同一の場合、`<-entry@morpheme/canonical` は省略される。また、対応する形態素オブジェクトが存在しない場合、両者とも省略される。

`<-entry@phrase` は句オブジェクトから見出しオブジェクトに張られたリンクによって生成される逆関係素性である。対応する句オブジェクトが存在しない場合、省略される。

`<-entry@sentence` は文オブジェクトから見出しオブジェクトに張られたリンクによって生成される逆関係素性である。対応する文オブジェクトが存在しない場合、省略される。

5.1.3 形態素オブジェクト

表 3 に形態素オブジェクトの素性一覧を示す。

素性名	値の内容
`=id`	ID
`=name`	名前
`->entry@morpheme`	形態素の表現形
`->entry@morpheme /canonical`	形態素の正規形
`->word-class`	品詞（語彙カテゴリー）
`<-morphemes`	この形態素を含む文の集合
`ja-form`	日本語表記
`ja-kana`	日本語でのよみ
`ja-conjugation-type`	日本語での活用の種類

表 3: 形態素オブジェクトの素性

[2] 複数文字からなる場合にも、なんらかの素性を付与すべきかも知れない。

->entry@morpheme(/*canonical*) の値は見出しオブジェクトの集合である。サブドメイン canonical は正規形を示し、これがないものは表現形を示す。表現型と正規形が同一の場合、->entry@morpheme/canonical は省略される。

->word-class の値は品詞オブジェクトの集合である。

形態素オブジェクトには文オブジェクトから ->morphemes が張られるために、その逆関係素性 <-morphemes が生成される。これはこの形態素オブジェクトを要素として持つ文オブジェクトの集合を値として持つ。

ja-form, ja-kana, ja-conjugation-type は MeCab 用古典中国語形態素コーパスや辞書との間でラウンドトリップ変換を可能にするために設けているもので、それぞれ文字列を格納している。

=name は形態素オブジェクトのユニークな名前[3]を示すためのものであり、オブジェクトの表示にも用いられる。この素性値は MeCab 用古典中国語形態素コーパス各フィールドの情報を

> 表現形 “ ” “(” 正規形 “)” “ ” “[” 大品詞 “,” 品詞 “,” 意味素性$_1$ “,” 意味素性$_2$ “]” “ ” “(” 日本語表記 “ ” “(” よみ “)” “,” 日本語での活用の種類 “)”

のように並べて作っている。また、コメントが存在する場合は、この後ろにタブと “;” を置きその後にコメントを続ける。

ID を示す ID 素性 =id には =name の値を変換したものを用いている。

5.1.4 文オブジェクト

表 4 に文オブジェクトの素性一覧を示す。

形態素コーパスの要素となる文は形態素オブジェクトの列からなるが、これを ->morphemes で表現する。また、形態素解析の揺れに対応するために、文の文字列オブジェクトを別に設け、->entry@sentence で表現する。この値は見出しオブジェクトの集合である。

素性名	値の内容
=id	ID
=name	名前
->entry@sentence	文の見出し文字列
->morphemes	形態素の列
source/file-name	コーパスのファイル名
source/sentence-number	コーパス内での文番号

表 4: 文オブジェクトの素性

また、文オブジェクトの集合から元のコーパス・ファイルが復元できるようにするため、source/file-name にコーパスのファイル名、source/sentence-number にそこでの文番号を格納している。

文オブジェクトは source/file-name の値で示されるコーパスにおける『文の出現』を示すものなので、その ID を示す素性 =id の値はコーパスのファイル名と文番号を “/” でつないだものにしている。また、文オブジェクトの名前 =name の値には

> 文の見出し文字列 “(“ 文の ID “)”

というものを用いている。

5.2 リンクを用いた揺れの表現

4 節で述べたような『メタ修正』を含むような効率的修正を実現するためには、内容を直接修正するよりも、修正（揺れ）を示すリンクをオブジェクト間に張るのが良いといえるが、これは形態素オブジェクト間や品詞オブジェクト間のリンクで実現できる。

例えば、「幼」(幼い) は、旧体系では「v, 形容詞」という品詞にしていたが、新体系では「v, 動詞, 描写, 形質」という品詞にしているが、この時、旧体系の形態素オブジェクト「幼 [v, 形容詞]」に関係素性 <-obsoleted を付与し、その値に新体系の形態素オブジェクト「幼 [v, 動詞, 描写, 形質]」を入れることで、形態素オブジェクト「幼 [v, 動詞, 描写, 形質]」と形態素オブジェクト「幼 [v, 形容詞]」の間に ->obsoleted というリンクが張られ、両者が新旧の関係にあるこ

[3]同じ名前を持つ他の形態素オブジェクトが存在しないことが保証されている。なお、他の genre には同じ名前のオブジェクトも存在可能であるが、genre が異なるために、違うオブジェクトとして認識される。

とが表現できる。また、これを一般化して、形態素オブジェクト間だけでなく、品詞オブジェクト間にも同様のリンクを張ることができる。この時、旧形式の品詞オブジェクトから新形式の品詞オブジェクトへの n 対 1 になっている場合にこの一般化を採用するというヒューリスティクスを用いることで、手作業でのチェック対象を減らすことができると考えられる。

6 おわりに

MeCab 用の古典中国語形態素コーパスの Concord/EgT を用いた Linked Data 化を試みた。これは現在「CHISE IDS 漢字検索」(`http://www.chise.org/ids-find` の検索結果の各行の先頭にリンクされた CHISE-Wiki の各頁で試験的に公開している。

古典中国語形態素コーパスを形態素見出し文字列、品詞、形態素、文の文字列、(コーパス中の) 文という 5 種類のオブジェクト間のグラフとして構成することで、品詞 (品詞体系) の揺れや文の形態素解析の揺れ等を簡単に抽出できるようになった。また、揺れている複数のオブジェクトに対して、変換元と変換先を示すリンク (関係素性) を張ることで、効率的な正規化が可能になると考えられる。また、形態素見出し文字列を介して CHISE 文字オントロジーとリンクしたことで、文字と形態素、形態素列、テキストをつないだ Linked Data が構成できた。これは、漢字の用例や意味を理解する上でも有用であるといえ、文字処理と形態素解析を併用したより適切な漢字処理を実現する上でも重要な基盤になり得るものだと考えられる。

参考文献

[1] Ya-Min Chou and Chu-Ren Huang. Hantology: An ontology based on conventionalized conceptualization. In *Natural Language Processing — Second International Joint Conference (IJCNLP 2005), Jeju Island, Korea, October 11-13, 2005, Proceedings*, 2005 年 10 月.

[2] Ya-Min Chou and Chu-Ren Huang. Hantology-a linguistic resource for chinese language processing and studying. In *LREC 2006 (5th edition of the International Conference on Language Resources and Evaluation*, 2006 年 5 月.

[3] International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*, 2011 年 3 月. ISO/IEC 10646:2011.

[4] Taku Kudo, et al. MeCab (和布蕪): Yet Another Part-of-Speech and Morphological Analyzer. `http://mecab.sourceforge.net/`.

[5] Tomohiko Morioka. CHISE: Character processing based on character ontology. In *Large-scale Knowledge Resources (LKR2008)*, No. 4938 in LNAI, pp. 148–162, 2008 年 3 月.

[6] 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 抄録集, Vol. 4, , 2006 年.

[7] 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, 第 2010 巻, pp. 33–40. 情報処理学会, 情報処理学会, 2010 年 12 月.

[8] 守岡知彦. CHISE の階層的素性名の RDF 化の試みについて. 情処研報, Vol. 2013-CH-97, No. 3, pp. 1–6, 2013 年 1 月.

[9] 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, 第 2012 巻, pp. 39–46. 情報処理学会, 情報処理学会, 2012 年 11 月.