中期朝鮮語文献の電子データ構築に関するいくつかの問題
―XMLの利用を中心に―

須賀井 義教

1．はじめに

本稿は、中期朝鮮語[1)]、特に15世紀のハングル文献資料について、その電子データ化の方法を検討するものである。本稿では特に、Extensible Markup Language（拡張可能マーク付け言語、XML）を利用して文献資料を電子データ化する際、どのような問題点が考えられるかを整理する。

言語の歴史的変化を明らかにする上で、文献資料が重要であることは言を俟たない。また近年にはコンピュータ技術の発展により、これら様々な文献資料を電子テキストとしてデータ化し、コーパスを構築して研究に活用することが容易となった。朝鮮語の文献資料についても、大韓民国の国立国語院などを中心に電子データ化が進められているが、その形式は単純なプレーンテキスト形式であるか、あるいは特定のソフトウェアに依拠した形式での入力、配布となっている。また、文献内のさまざまな構造が反映されておらず、検索などの処理において困難を感ずることが少なくない。本稿では、これらの不便を克服し、朝鮮語研究に有用な電子データの作成を目指すという観点から、まず資料を電子データ化する際の問題点について整理し、その解決策を検討する。

朝鮮語の様々な文献資料のうち、本稿では特に15世紀のハングル文献について扱う。15世紀半ばのハングル創製により、朝鮮語の姿を初めて完全な形でとらえることができるようになった[2)]。近年にはそれ以前の朝鮮語についても多くの研究がなされているが、やはり15世紀ハングル文献の持つ重要性に変わりはない。古代朝鮮語の研究も、結局は15世紀の言語を拠りどころとして行なわれているためである。この点で、15世紀のハングル文献を電子データ化することは、朝鮮語史研究において重要な役割を果たすといえよう。電子データとして蓄積しておくことで、個々の研究者が検索・参照を容易に行なうことができ、また計量的な研究が可能となる。

さて前述の通り、これまで提供されてきた朝鮮語文献の電子化テキストは、文献の内部構造をそれほど重視せず、単純なプレーンテキストの形で入力されたり、韓国内で主に流通するワープロソフトウェアの文書形式により配布されていた。こうした点で、これまで

の電子データはその検索や再利用が柔軟に行なえず、特定の環境（ワープロソフト、オペレーティングシステムなど）でのみ利用するためのものであった。以下、これらの電子データについて簡単に触れておく。

まず、プレーンテキストの形式で入力された朝鮮語文献の電子データは、朝鮮語史研究者の間で流通しているもので、その形式は概ね以下の通りとなっている：

例1：プレーンテキスト形式の朝鮮語文献電子データ（一部）

<釋詳 6:15a> 護彌 닐오ᄃᆡ 소리ᄲᅮᆫ 듣노라 婆羅門이〔中略〕舍衛國으로 가리 잇더니
<釋詳 6:15b> 婆羅門이 ᄀᆞᆯ왈ᄒᆞ야 須達의손ᄃᆡ 보내야ᄂᆞᆯ〔後略〕

各行の始めには文献名の略号と巻数、コロンに続いてページとその表裏が、「< >」に入れて表示されている。例1の「<釋詳6:15a>」は、『釈譜詳節』巻六の第15張表を指す。末尾のローマ字が「b」の場合には張次の裏を指している。文献の各ページを1行に入力し、行頭にマークをつけた形である。本文中に現れる注釈（割注）部分は「[ ]」でくくって示している。ファイル自体はUnicode[3]を用いて記述されている。

この電子ファイルの利点は、15世紀から20世紀に至るまで、影印などが手に入る文献についてはほとんど網羅されているという点であろう。様々な検索プログラムを利用して、例えば15世紀の文献のみ、あるいは17世紀以降の資料のみを検索するといったことが可能である。

しかし、分かち書きや記号の使用など入力の方式が入力者によって一定しておらず、ファイルの使用に一定の注意が必要である。また分かち書きと関連して、ある文節がページをまたぐ場合、文節単位で前の行に送り込んでいるという問題がある。例えば例1の1行目末尾に「舍衛國으로 가리 잇더니」（舎衛国に行くものがいたが）とあるが、原資料では「잇더니」の「잇」と「더니」の間でページが変わっている。原資料におけるページ区切りを示すマーカーがないため、原本の体裁を復元するということが不可能である。他に、諺解[4]部分だけでなく、漢文や漢字音まで入力されているファイルもあるが、諺解部分と漢文部分とが区別なく羅列されており、諺解部分のみ検索することが難しい。また、これらの電子ファイルにおいては中期朝鮮語に存在した「声調」[5]を示す傍点が入力されておらず、データとしては内容が十分でないといえよう[6]。

また記述に際し、文書構造をマークアップするTEI（Text Encoding Initiative）[7]ガイドラインを用いる場合（3.1節で後述）もあったが、この構造マークアップを積極的に活用

した事例は、管見の限り多くはない。例えば韓国における韓国語情報化プロジェクト「21世紀世宗計画」ホームページでは、朝鮮語史の文献資料から用例を検索することができるが、文献に関する詳細情報を表示する際、著者や入力情報など、TEIマークアップによる情報が示される。ただし、同じく21世紀世宗計画が作成、公開しているコーパス「세종말뭉치」については、同時に配布されている検索プログラム「글잡이」でこのマークアップを活用しているわけではないようである。

これら既存の電子データを利用する際に生じる不便を克服するため、本稿では以下の点を念頭に置いて、中期朝鮮語文献をデータ化する方法を検討したい。即ち、①文献本来の体裁をおおよそ復元できること、②形態の検索や索引の作成など、言語研究に応用しやすい形式を備えていること、③利用に特殊なアプリケーションを必要としないこと、以上の3点である。①は主に文献の表示、②は研究への活用、③は利用に当たっての利便性を念頭に置いたものである。これらの留意点を満たすデータ化の方法として、本稿ではXMLを用いたデータの記述を提案する。

なお、言語研究におけるXMLの活用事例として、日本国内の研究では国立国語研究所編（2005）が挙げられる。近代日本語の資料である雑誌『太陽』をXMLにより電子データ化するだけでなく、コーパスとして利用するための検索、形式変換などのツール開発（小木曽智信 2005、山口昌也 2005など）も行なっており、大いに参考となる。

また、本稿の筆者は既に、いくつかの中期朝鮮語文献をXMLにより記述し、その電子データをインターネット上で公開している[8)]。これについては後で詳しく述べるが、公開当初の目的は、中期朝鮮語の文献についてどこからでも参照できるようにすることであった。さまざまな再利用が可能であると考え、まずはXMLを利用した。本稿では、これらの作業から一歩進んで、言語研究の資料として活用できる電子データの構築について検討してみたい。

2．XMLについて

2.1　XMLの概要

電子データ記述の方法を検討するのに先立ち、まず本稿で利用する「XML」について簡単に説明しておく。

XML（Extensible Markup Language、拡張可能マーク付け言語[9)]）は、電子化された文書の交換や多目的利用のために考案された、SGML（Standard Generalized Markup Language、標準一般化マーク付け言語）のサブセットである。インターネットのホーム

ページ記述に用いられるHTML[10]もSGMLに基づいたコンピュータ言語であるが、HTMLは文書の表示を主な目的としており、マーク付けのためのタグ（tag）が既に決められている。それに対し、文書の構造記述を主眼とするXMLではタグを自由に定義できるのが特徴である。開始タグ（「<タグ名>」の形式）と終了タグ（「</タグ名>」の形式）とで囲み、要素を記述する。それぞれの要素は、さらに別の要素を入れ子にすることもできる。その他技術的な詳細については、芝野耕司（2000）やWorld Wide Web Consortium（W3C）のホームページなど、インターネット上の情報を参照されたい[11]。

2.2 なぜXMLを用いるか

さて、本稿でXMLを利用する理由として、①文書を構造化して記述できる、②データの加工、再利用が容易である、③文字符号化方式（文字コード）としてUnicodeを採用している、などといった点が挙げられる。

まず①について、中期朝鮮語に限らず、文献資料はその内部にさまざまな構造を持っている。例えば表題や目次など、「本文」として扱われない部分であるとか、あるいは本文の中でも会話部分であったり注釈部分であったりといった具合に、単に文字が並べてあるというわけではない。文献の物理的構造について言えば、欠損や落丁、また欄外の書き込みや校正の痕跡など、いろいろな要素が存在する。先の例1に挙げたようなデータ記述は、これらの構造のうちごく一部、本文と注釈の区別などを反映するだけであり、その他さまざまな構造についての情報を盛り込むことができない。例えば、地の文ではなく会話部分に現れる形式について検索したい、注釈部分は除いて語句を検索したいなどといった場合、例1のようなデータではその実行が困難である。それに対し、XMLを利用し、文献の構造を定義したタグを用いることで、こうした文献の構造を記述することが可能となる。また、文献の内容だけでなく、刊年や著者など、文献に関する付加的な情報を盛り込むことができる。こうしたXMLの利点を活かし、文献のデータ化を目指したものが前述のTEIである。しかし、TEIではさまざまなタイプの文献を対象としており、その体系は膨大なものになっている[12]。よって本稿ではまず、中期朝鮮語文献に特化した電子データ化の枠組みを考える。

②については、XSLT[13]といった技術やPerlなどのテキスト処理プログラム言語を用いて、さまざまな形式に加工することができるという点が挙げられる。ウェブブラウザなどを用いて表示するよう、HTMLなどへ変換したり、必要な部分だけ選択して検索したりといったことも可能である。

さらに、中期朝鮮語の文献を電子データとして記述するにあたって、最も重要なのが③であるといえよう。中期朝鮮語には、現代では使われない文字や字形が多く含まれている。また近代の朝鮮語文献には日本語や中国語の教科書などもあり、その表示にはハングルと漢字だけでこと足りるものではない。従来は中期朝鮮語の入力や表示の際、「アレアハングル」など特定のアプリケーションが必要であった。しかしUnicodeでは、これらのさまざまな文字集合をひとつの枠組みで扱うことができ、テキストエディタがあればファイルを閲覧することができる。アプリケーションに依存しない[14]という点で、Unicodeをサポートする XML の利用価値は高いと考える。

3．XMLによって記述を行なう際の問題点

では、上述したXMLを中期朝鮮語文献の電子データ化に用いる際、どのような問題が考えられるか、以下で整理してみる。本稿では、構造化の単位に関する問題、記号や入力できない文字の処理について検討する。

3.1　構造化の基準となる単位

豊島正之（2001）、近藤泰弘（2003、2004）などでも取り上げられているが、文献資料を電子データとして記述する際、どのように構造化を行なうかという問題がある。例えば近藤泰弘（2003：72）では、日本の古典語コーパスに含まれる文書構造として以下の4つを挙げている：

1　作品の論理構造（巻・章・個々の和歌（その属性としての歌番号・作者名）・章節名）
2　原写本・刊本の物理的構造や表記（冊・丁数・表・裏・行数等）
3　翻字印刷された活字本の物理的構造（冊・ページ数・行数等）
4　文書中に現れる様々な諸要素（傍訓・ルビ・小書き・割り注・虫食・字種等）

これらの構造のうち、どれを基本のブロックとしてデータ記述を行なうかという点が大きな問題であるが、これは中期朝鮮語文献の場合も同様である。筆者が作成した中期朝鮮語の電子データでは、上記のうち2に該当する、原資料の物理的構造や表記を生かした方法を採用した。以下に例を見てみよう（各行左端に行番号を付した）：

例２：筆者が作成した中期朝鮮語電子データの例（『阿弥陀経諺解』活字本、1461 年？）

```
<?xml version='1.0'?>
<book name="阿弥陀経(活字本)">
<title>佛說阿彌陁經</title>
<info>
    <source label='出典'>韓國書誌學會(1993) "季刊 書誌學報 第 10 號"</source>
    <input label='入力者'>SUGAI, Yoshinori</input>
    <date label='修正日'>2003/08/20 21:14</date>
</info>
<text>
    <page no="01" side="a">
        <line no="01" type="t">佛說阿彌陁經</line>
    </page>
    <page no="01" side="b">
        <line no="07" type="t">·이 ·곧·호·물 ·내 듣ᄌᆞ·ᄫᆞ·니 ᄒᆞᆫ·ᄢᅴ 부:톄 舍衛國 祇^</line>
        <line no="08" type="t">樹 給孤獨園·에 :겨·샤 ·큰 比丘 :즁</line>
        <line no="09" type="t">千二百 :쉰 :사ᄅᆞᆷ·과 ᄒᆞᆫ ·ᄃᆡ 잇·더·시니</line>
    </page>
    <page no="02" side="b">
        <line no="07" type="t">:다 大阿羅漢·앳
            <note>[大·ᄂᆞᆫ ·클 ·씨·라. 阿羅漢·은 殺賊·이·라 ·혼</note>
        </line>
        <line no="08" type="t">
            <note>·ᄠᅳ디·니 殺·은 주·길 ·씨·니 〔中略〕 나·디 아·니=·탓 ·ᄠᅳ디·니</note>
        </line>
        <line no="09" type="t">
            <note>ᄂᆞ외·야 生死ㅅ 果報·애 〔中略〕 應·은 ·맛당ᄒᆞᆯ ·씨·니 ㅅ</note>^
        </line>
    </page>
〔以下略〕
```

例２では、「book」というルート要素を最上位として、その下位に「title」（文献の表題、ここでは内題）、「info」（文献に関する情報）、「text」（本文）という３つの要素を配置したものである。「book」要素の属性（attribute）として「name」属性を与えているが、これは文献の略称を指す。また、「info」要素は「source」（出典）、「input」（入力者）、「date」（入力あるいは修正の日付）といった要素を含んでいるが、これらの要素につい

ては再考の余地があるだろう。本文（「text」要素）は複数の「page」要素からなり、また「page」要素は複数の行、つまり「line」要素を含む。「page」要素の「no」属性は張次の番号を、「side」属性は表（a）、裏（b）を表している。「line」要素も同様であるが、「type」属性は本文（t）と注釈（n）を示す。「note」要素は注釈部分を指している。本文と注釈を示す記号や、「=」や「^」といった記号については、主に趙義成（2000、2005）や조의성（2002）を参考にした。また、「・」や「：」といった記号により、声調の傍点を示した。このようにして作成したXMLファイルを、XSLスタイルシートによりウェブブラウザで表示できるようにした例が、以下に示す図1である：

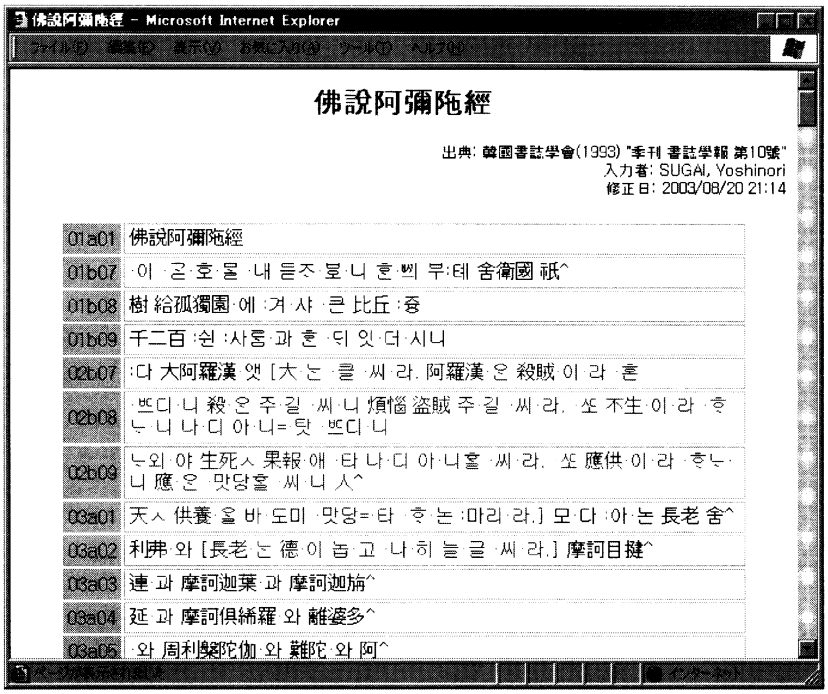

図1：ウェブブラウザにおける「阿弥陀経諺解」ファイルの表示例

ここでは「title」要素を見出しとして配置し、「info」要素に含まれる内容を付加情報として表示した。「text」以下、本文の内容については行ごとに枠でくくり、「張次」「表裏」「行数」を左端に表示してある[15]。また、注釈部分は色を変えて表示するようにした。

この電子データは、文献の物理的構造をそのままタグで置き換えて表現したものであり、例2のようなデータは、文献の体裁にあわせた表示には向いているが、その内容を検索したりする場合には、文の単位でテクストを切り出したりするなどの加工作業が必要となる。特に、例2の20行目、23行目、26行目を見ると、全てひと続きの注釈部分である

にもかかわらず、複数行にまたがっているために、別個の「note」要素となってしまっている。これは、要素は必ず入れ子になっていなければならないというXMLの制約によるものである。この細切れになった部分をどうやってつなげるか、検討する必要があるだろう。

それに対し、「文」や「章」といった文献内容の論理構造を基準として記述した場合、どのような問題があるだろうか。論理構造を基本とした構造化の例として、TEIによる中期朝鮮語文献のマークアップ（21世紀世宗計画ホームページにて配布）や、日本の「太陽コーパス」（国立国語研究所編 2005）などがある。ただし、いずれも文や形態素といったレベルまで細分化して構造化しているわけではない。

まず、TEIによる中期朝鮮語文献のマークアップの例を見てみよう。『蒙山和尚法語略録諺解』（1472年刊）を入力したファイルから、一部を抜粋する：

例3：TEIによるマークアップの例（『蒙山和尚法語略録諺解』）

```
1:  <pb n='蒙山 13a'>
2:  <p>覺圓上座ᄂᆞᆫ 아ᄂᆞᆫ다 모ᄅᆞᄂᆞᆫ다 微妙ᄒᆞᆫ 아로미 圓滿히 ᄇᆞᆯ마란ᄃᆡ 반ᄃᆞ기 趙州ㅣ 엇던
3:  面目인 ᄃᆞᆯ 아로리라 無ᄀᆞ 字ᄅᆞᆯ 닐온 ᄠᅳ든 엇뎨어뇨</p>
4:  <p>구믈구믈ᄒᆞᄂᆞᆫ 衆生이 다 佛性<pb n='蒙山 13b'>이</p>
5:  <p>잇거시니 趙州ᄂᆞᆫ 어ᄃᆡᄅᆞᆯ 因ᄒᆞ야 업다 니ᄅᆞ뇨</p>
```

例3では、「pb」（page break）と「p」（paragraph）の2つのタグが使われている。入力されているのは漢文を除いた諺解部分であり、例3に示したのは第13張の表と裏にまたがった部分である。元の文献は以下の通りとなっている[16]：

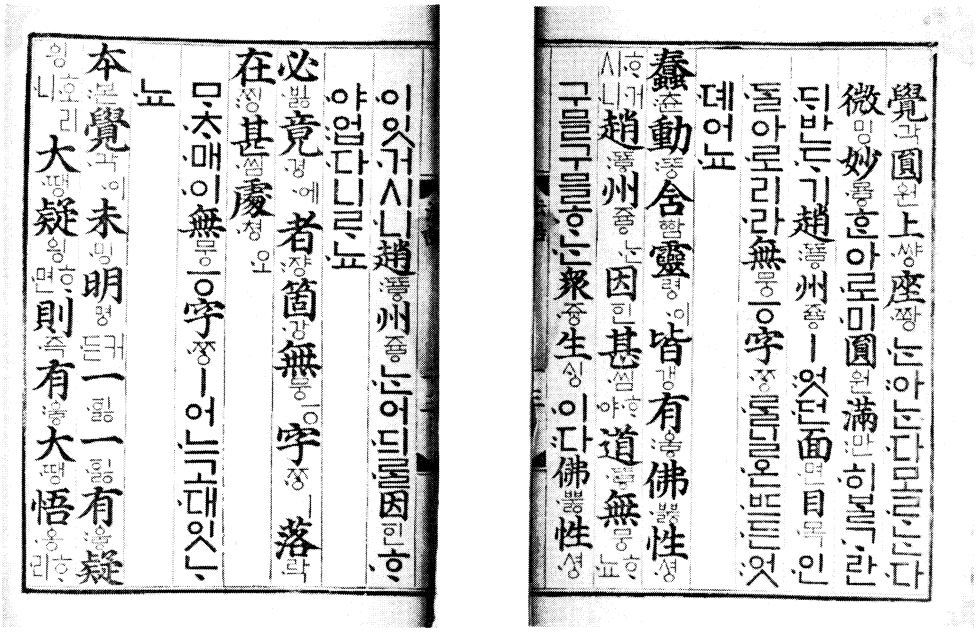

図２：『蒙山和尚法語略録諺解』第13張表（右図）と裏（左図）

上の図２で、それぞれ１字下げで始まっている段落が諺解部分である。例３において「p」タグで囲まれているのは、基本的にこの１字下げの部分に該当する。この「p」タグで囲まれた部分には、１つ以上の文が含まれている。ただし、13張の表から裏へまたがる「衆生이 다 佛性이 잇거시니」（衆生がみな仏性があるが）という部分（例３の４行目から５行目）では、文の途中であるにもかかわらず、「佛性이（仏性が）」の後ろで「p」タグが閉じられている。結局、何を基準として「p」タグを使用しているのかという点が曖昧である。また、文の終わりを示すマーカーもない。検索や索引作成など何らかの処理を行なう際に、このままでは文を単位として処理することもできず、また段落を単位として処理すると文の途中で切れてしまうということになってしまう。この問題はつまるところ、データをどのように活用するかといった明確な目的がなく、データの処理方法まで考えずにデータ化を行なっているためと見受けられる。データ化の枠組みに、ある程度の汎用性を持たせる必要はあるだろうが、どのように処理するかということを念頭において、データ記述の枠組みを設定しなければならないと考える。

さて、文献内容の論理構造を基準としてデータ記述したもう１つの例として、「太陽コーパス」を挙げた。その設計概要（田中牧郎 2005：31）では「言語の階層構造に即した厳密な構造化は行わず、文の認定も行わない」としている。しかしコンピュータで処理する場合などに備えて、「段落や文の認定に代わる便宜的な切れ目を設けて構造化を行っ

た」という。例を見てみよう：

例4：「太陽コーパス」のデータ記述例（1895年1号、「太陽の発刊」より）

```
<s>歸朝の後、</s>
<s>彼邦に譲らざるべき大雑誌を發行せんと<l 位置="P001D05" />計畫したるも、</s>
<s>時既に歳晩に迫りて之れを決行するの暇なく、</s>
<s><注 原文="已む" 分類="A 誤字通用">已む</注>を得ず本年を期したりしが、</s>
<s>昨年六月日清開戰の事起りしより、</s>
<s>事躰頗る大なるを以て日清戰爭實記<l 位置="P001D06" />を發行せり。</s>
```

例4に見えるとおり、「。」または「、」で区切られた範囲を「擬似的な文」（田中牧郎2005：31）として「s」タグでマークアップしている。これにより、厳密には「文」ではないが、ある程度の区切れを単位として処理を行なうことが可能となっている。実際に、「太陽コーパス」から検索を行なうアプリケーション「たんぽぽ」は、この「s」要素を1つずつ読み込んでそれぞれに検索処理を行なっている（小木曽智信 2005：103）。

また、例4では『太陽』原文での位置を示すのに、空要素「l」とその属性「位置」を用いている。原文テキスト内でのページ区切りや位置を示すこのような方法は、先のTEIによるマークアップ「pb」（page break）と同様である。

以上、文献の物理的構造を基準としたデータ記述、そして文献内容の論理構造を基準としたデータ記述について見てきた。いずれの場合にも長所と短所があると思われるが、検索や索引作成の便を考えれば、後者の方法がより利用しやすいのではないかと考えられる。検索や索引の作成は、結局のところ言語上の単位を対象として行なうのであって、資料におけるページ区切りなどといった物理的な構造は、優先度がより低いといえよう。もちろん、出現箇所を示すために張次や行番号などは必要である。そのため、「太陽コーパス」のように「文」（あるいは「擬似的な文」）を基本単位として、その中間にページ区切りを挿入していく方法が有効ではないかと考える。特に、厳密に「文」という区切りを設定するのではなく、ある程度ゆるやかな枠組みをもって、その単位を設定していくのが望ましいだろう[17]。

なお、文よりも下位の単位として「文節」などを設定するかどうかという点についても検討すべきであるが、本稿では扱わず、今後の課題としたい。

3.2 傍点や入力できない文字などの扱い

3.2.1 傍点の処理

先に中期朝鮮語文献の例として、図2に図版を掲げた。それを見れば分かるように、それぞれのハングルの左側に、声調を表す傍点（註5を参照）がついている。これをデータとしてどのように記述するかという問題がある。本稿では今のところこの問題に対する答えを持っていないが、2つの方法を考えうる。

まず、文献の電子データ自体を記述する際に、それぞれの文字ごとにタグをつけ、その属性に傍点情報を与えるという方法である。例えば図2の第13張表、4行目から5行目にかけて「:엇•뎨어•뇨」（どうしてか？）という部分がある。「:엇」が上声、「•뎨」「•뇨」がそれぞれ去声、傍点のない「어」は平声を示す。この「:엇•뎨어•뇨」をデータとして記述する際に、以下のように行なうことができよう：

例5：データ記述における傍点の処理例

```
<t pitch="r">엇</t><t pitch="h">뎨</t>어<t pitch="h">뇨</t>
```

ここでは「t」(tone) というタグを設定し、その属性「pitch」に上声（上昇調）であれば「r」(rising)、去声（高調）であれば「h」(high) を記述する。平声は「t」タグで囲まない。しかし、ひとつの文節が細切れになってファイルの可読性があまり高くなく、入力の手間もかかるため、この方法はさらに検討してみる必要があるだろう。

もうひとつの方法は、「:엇•뎨어•뇨」とそのまま傍点をつけて記述しておき、検索や索引作成の処理を行なう段階でこれらの傍点を無視するというものである。いずれにせよ、どのような検索を行なうか、またどのような処理を行なうかという点を明確にし、その上で記述の方法を検討すべきであるだろう。

3.2.2 入力できない文字の表示などについて

また、コンピュータ上で入力できない文字、あるいは表示できない文字などがある。特に漢字の入力・表示において、フォントに存在しないケースなどがあり、問題が生じる。例として、『阿弥陀経諺解』（活字本）の第3張表9行目、「阿■樓馱」（阿那律の別称[18]）という部分を挙げておく。この■の部分は、「少」と「免」を上下に組み合わせた漢字が該当する。筆者が既に公開している電子データでは、「今昔文字鏡」[19]の文字番号により記述することとした。先の■に該当する漢字は「今昔文字鏡」の番号で「007511」が割り当

てられている。この番号を利用して、表示する際にGIFイメージを埋め込むこともできるだろうが、公開しているファイルでは「阿（亍=007511）樓馱」のように、漢字音と文字鏡番号とをカッコでくくって表示するのみとした。

「今昔文字鏡」に収録されている場合は上記のように処理することができるが、未収録の漢字や記号の場合、処理が困難である。偏や旁の組み合わせなどをカッコに入れて表示する方法などが考えうるが、検索や表示の際にどう処理するか、検討すべき課題である。

4．おわりに

本稿では、中期朝鮮語の文献を電子文書として記述する際に、どのような問題が考えられるか検討した。本稿で扱ったのは、データ記述の際に構造化の基準とする単位の問題、傍点や入力できない漢字の処理についての問題である。ここではそれぞれの問題についていくつかの例を整理したのみで、データ記述の方法を明確に打ち出すことができなかった。

他に、本稿で扱わなかった問題として、形態素解析に関連する問題、虫食いや落丁などによる欠損についての問題などがある。

中期朝鮮語では表記の様相がさまざまに現われ、特定の形態を検索するのに困難を感じることが少なくない。この負担を軽減するためにも、形態素解析を行なって、その結果を検索に利用することができるのが望ましい。しかし、個別の形態素自体をどのように規定するか、基本となる形態をどのように設定するかという問題もあり、今後さらに検討しなければならないだろう。

また、欠損により文献の一部が欠落している場合、どのような処理を行なうかという問題がある。語の中間、あるいは末尾が欠落している場合、その部分を補うのか。落丁がある場合には、それを示す空要素タグを挿入するべきか。

これらの課題について今後さらに検討し、いくつかの文献を実際に電子データ化して、データ記述のための枠組みを構築していきたいと考える。

注

1） 朝鮮語史の時代区分について、本稿では河野六郎（1955：428）による以下の区分に従う：
古代朝鮮語　諺文発明（1443年）以前
中期朝鮮語　1443年より1592年の壬辰の役まで（15世紀中葉から16世紀末まで）
近世朝鮮語　それ以降現代まで

2) ハングル創製以前には、漢字の音や訓を用いたいわゆる「借字表記」が行なわれていた。借字表記をはじめとする朝鮮語史の概要については、李基文（1972：1975）などを参照のこと。

3) 世界のさまざまな文字を一つの文字コードに収めるという目標のもとに、Unicode Consortiumを中心に開発が進められている文字コード体系。詳細はUnicode Consortiumのホームページ、安岡孝一・安岡素子（1999）などを参照のこと。

4) 15世紀半ばのハングル創製以降、ハングルは主に仏教経典などの翻訳に用いられた。それらの文献は通常、経典の原文を漢文で示し、それにハングルによる翻訳、即ち「諺解」を付すという形式となっている。

5) 「声調」といっても中国語におけるそれとは異なり、高低のアクセントを示すものであったといわれる。低調（無点）、高調（一点）、上昇調（二点）のように、文字の横に点をつけることで示した。

6) この点については、既に趙義成（2000）の指摘がある。

7) 人文科学の文献資料などを電子化するための標準策定を目指すプロジェクト。「言語データをSGMLに基づいてテキストとして流通させる為の、文書諸要素（element）の定義集」（豊島正之 1994：2000：2）。XMLアプリケーションの一つでもある。概要についてはTEIホームページ、강범모（2003）などを参照のこと。

8) 「MEMORANDUM - 資料室」（http://porocise.hp.infoseek.co.jp/archive/）を参照。

9) 以下、XML関連の用語については、芝野耕司（2000）を参考にした。

10) Hyper Text Markup Languageのこと。インターネット上の文書を記述するための言語であり、複数の文書をむすびつけるハイパーリンクや、画像、表の表示など、様々な表現力を持っている。

11) 言語コーパス構築におけるXMLのタグ付けの例としては、国立国語研究所の言語コーパス整備計画「KOTONOHA」ホームページ内「電子化形式」が分かりやすい。

12) 単純に分量で計算できるわけではないが、TEI Consortiumで配布している最新の第5版ガイドラインは、1300ページを超える大部なものである。

13) XSL Transformations、XSL変換とも。XML文書の構造を変換するための言語である。

14) ただし、アプリケーションには依存しないものの、その表示には中期朝鮮語で用いられる字形、いわゆる「古ハングル（옛한글）」を収録したフォントが必要である。筆者ホームページの「中期朝鮮語とUnicode」（http://porocise.hp.infoseek.co.jp/memo/mk_uni.html）を参照。

15) 例えば「02b09」であれば、第2張の裏、9行目を指している。

16) 図2の画像は、韓国の国立国語院が運営する「디지털 한글박물관」（デジタルハングル博物館）ホームページよりダウンロードしたものである。

17）文の認定と関連して、中期朝鮮語において終止形語尾を基準に「文」を認定すると、ひとつの文が極端に長くなってしまう場合がある。また、李賢熙（1994）で言及されているように、一部の接続形語尾は意味段落を成すことがある。よって、終止形語尾の出現をもって「文」とするよりも、終止形語尾と一部の接続形語尾とを「擬似的な文」の区切りとするのが適当ではないかと考える。

18）『総合佛教大辞典』による。

19）「今昔文字鏡」は、10万字を超える漢字を収録したフォントパッケージ。文字鏡研究会のホームページを参照のこと。「フォントセンター」で収録文字の検索などを行なうことができる（http://www.mojikyo.org/html/fontcenter/giflink.html）。

参考文献

小木曽智信（2005）「構造化テキストを直接利用するアプリケーション～『プリズム』と『たんぽぽ』～」（国立国語研究所編 2005 に収録）.

河野六郎（1955）「朝鮮語」,『世界言語概説　下巻』, 東京：研究社, pp.357-439.

国立国語研究所編（2005）『雑誌『太陽』による確立期現代語の研究――『太陽コーパス』研究論文集』, 東京：博文館新社.

近藤泰弘（2003）「古典語のコーパス」,『日本語学』4月臨時増刊号, 東京：明治書院, pp.62-81.

近藤泰弘（2004）「日本語コーパス言語学とコンピュータ処理」, 秋元実治他『コーパスに基づく言語研究――文法化を中心に』, 東京：ひつじ書房.

芝野耕司（2000）『SGML/XML が分かる本』, 東京：オーム社出版局.

総合佛教大辞典編集委員会（1987）『総合佛教大辞典』, 京都：法藏館.

田中牧郎（2005）「言語資料としての『太陽』の考察と『太陽コーパス』の設計」（国立国語研究所編 2005 に収録）.

趙義成（2000）「朝鮮語テクストのコンピュータ処理について――中期朝鮮語 KWIC 索引作成の場合――」,『県立新潟女子短期大学研究紀要』第 37 集, 新潟：県立新潟女子短期大学, pp.153-167.

趙義成（2005）『初刊本『釋譜詳節』統合 KWIC 索引　第 1 分冊　本文・正順索引』, 私家版.

豊島正之（1992；2000）「TEI からみた SGML のはなし」（『情報処理語学文学研究会会報』12 号, http://www.joao-roiz.jp/mtoyo/TEI/JALLC-12-TEI.pdf）

豊島正之（1994；2000）「TEI-P3 について」（『情報処理語学文学研究会会報』15 号, http://www.joao-roiz.jp/mtoyo/TEI/JALLC-12-TEI.pdf

豊島正之（2001）「XML の骨抜き利用法」（古典学の再構築―情報処理（A 0 3）班主宰研究集会, http://www.joao-roiz.jp/mtoyo/TEI/JALLC-TEIP3.pdf）

安岡孝一・安岡素子（1999）『文字コードの世界』, 東京：東京電機大学出版局.

山口昌也（2005）「構造化テキストに対応した全文検索システム『ひまわり』」（国立国語研究所編 2005 に収録）.

李基文（1972;1975）『韓国語の歴史』藤本幸夫訳，東京：大修館書店.

강범모（2003）“언어，컴퓨터，코퍼스 언어학”，서울：고려대학교 출판부.

李賢熙（1994）“中世國語 構文研究”，서울：新丘文化社.

조의성（2002）“月印釋譜（卷一）語彙索引”，서울：도서출판 박이정.

TEI Consortium（2008）*TEI P5: Guidelines for Electronic Text Encoding and Interchange*（http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf）.

〔参照ウェブサイト〕

TEI: Text Encoding Initiative　http://tei-c.org/

Unicode Consortium　http://www.unicode.org/

World Wide Web Consortium（W3C）　http://www.w3.org/

独立行政法人国立国語研究所「言語データベースとソフトウェア」
http://www.kokken.go.jp/lrc/

独立行政法人国立国語研究所「国立国語言語研究所のコーパス整備計画 KOTONOHA」
http://www.kokken.go.jp/kotonoha/

文字鏡研究会　http://www.mojikyo.org/

국립국어원（国立国語院・大韓民国）　http://www.korean.go.kr/

디지털 한글박물관（デジタルハングル博物館・大韓民国）
http://www.hangulmuseum.org/

21 세기 세종계획（21 世紀世宗計画・大韓民国）　http://www.sejong.or.kr/