共同研究プロジェクト「朝鮮語歴史言語学のための共有研究資源構築」
（平成 21 年度第 1 回研究会）

日時：平成 21 年 8 月 22 日（土）　14:00～17:00
場所：東京外国語大学アジア・アフリカ言語文化研究所 4 階研修室（405 室）

報告者名：李文淑（AA 研共同研究員，東京理科大学），須賀井義教（AA 研共同研究員，近畿大学）
報告タイトル：李文淑「中国黒龍江省尚志市で話される朝鮮語のアクセント」，須賀井義教「XML を利用した朝鮮語史資料の電子データ化」

報告内容：次ページより添付

中国黒竜江省尚志市で話される朝鮮語のアクセント

李文淑（AA 研共同研究員，東京理科大学）

本発表は、中国黒龍江省尚志市（サンジシ）で話されている朝鮮語のアクセントについて報告するものである。中国にいる朝鮮族の言語に関する報告は朝鮮族自治区に限られており、それ以外の地域で話される朝鮮語についての報告はほとんどない。今回は朝鮮半島の慶尚道から移住した移住 3 世代目のアクセントについて考察する。

名詞のアクセント体系は以下の通りである。表の横の数字は音節数を、縦の丸付きの数字はアクセント核の位置を示す。また、A は 1 音節卓立型を、B は下降が起こる直前の音節から高くなる型、即ち、高い音が 2 音節続く型を表す。ただし、⓪だけは最初から 2 音節目まで高い音が続くもので別途にし、=で表記する。ここで言うアクセント核とは音調の下降をもたらすもので、]で表記する。*が付いているところは単独形では語例が見つかってないが、助詞付きの形では現れるもので、単独形でも助詞付きでも現れないものは空欄にしている。

表 1.　アクセント体系

区分		1	2	3	4	5
	⓪	○=	○○=	○○○=	○○○○=	○○○○○=
A	①		○]○	*○]○○	*○]○○○	
	②	○	○○]	○○]○	○○]○○	○○]○○○
	③		○○]	○○○]	○○○]○	○○○]○○
	④				○○○○]	○○○○]○
B	③		○○]	○○○]	*○○○]○	○○○]○○
	④				○○○○]	○○○○]○
	⑤					○○○○○]

ここで特徴的な点は 1 音節語と 2 音節語である。1 音節語は単独形と 1 音節助詞を付けた場合、型の対立が現れず、全てが同じ形⓪で現れる。ここまでを見るとアクセントは弁別性がないように見えるが、2 音節助詞をつけると⓪と②で分かれる。ここで初めて型の特徴が現れる。

2 音節語は、語単独形だけを見ると第 1 音節は低く、第 2 音節だけが高い LH のパターンが三つ現れる。助詞をつけると、第 2 音節だけが高い②、第 3 音節だけが高い A-③、第 2 音節から 3 音節目まで高い音が続く B-③の三つで分かれる。特に、B-③は安定しておらず、A-③と一緒に現れるなどゆれが見られる。

次は、複合語アクセントについて述べるが、当言語の複合語アクセントは規則に従って現れる安定しているものではない。また、各要素の音節数とも関係しており、簡潔に整理できる規則ではないが、複合語に見られる一つの傾向として考えられる条件を規則として立てておく。X は前部要素のアクセント、Y は後部要素のアクセントを意味する。

表 2.　複合語アクセント規則

構造	性質・条件		複合語のアクセント
1+1	基本的に前部要素決定		X
1+2	語頭子音	濃音･激音･s･h	X
		平音	Y
	後部要素が○●③		Y
1+3	基本的には後部要素決定		Y(?)
2+1	前部要素決定		X
2+2	基本的には前部要素決定		X
	後部要素が○●③		Y
	前部要素が○●②		Y
2+3	基本的には前部要素決定		X
	前部要素が○●②		Y
3+1	前部要素決定		X
3+2	前部要素決定		X
3+3	前部要素決定		X

上の表 2 で見るように複合語のアクセントは語の構造と大きく関係している。つまり、複合語になった同じ 3 音節語でも 1+2 の構造と 2+1 の構造はアクセントが決まる条件が異なる。1+2 は後部要素決定型であるが、2+1 は前部要素決定型である。複合語全体で見ると前部要素によって決まるのが一般的であるが、限られた条件下では後部要素によって決まることもある。特に、後部要素が A-③の場合は、この A-③が保存されることが多い（1+2、2+2)。

また、B-③は単純語でもその特徴を保ちにくく、A-③の形と一緒に現れるなどゆれが見られたが、この B-③が前部要素になると後部要素によって複合語のアクセントが決まる。

朝鮮語史資料研究会（2009 年 8 月 22 日）【発表要旨】

XML を利用した朝鮮語史資料の電子データ化

――データ記述とツール作成の試み――

須賀井　義教（近畿大学）
sugaiy@kindai.ac.jp

1. はじめに

本稿は，朝鮮語史資料のうち，特に 15 世紀の朝鮮語資料を電子データとして記述する方法について検討し，その枠組みを構築することを目的とする．特に本稿では，XML（eXtensible Markup Language，拡張マーク付け言語）を利用した電子データの記述方法を提案する．

大韓民国の国立国語院などを中心として，朝鮮語史資料の電子データ化が既に進められているが，特定のソフトウェアに依拠したファイル形式が採用されていたり，文献内のさまざまな構造が反映されていないなど，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，朝鮮語史研究に有用な電子データを作成するという観点から，電子データ構築のための枠組み設計について検討していく．

本稿では，以下の諸点を念頭において，電子データ記述の枠組みを設定する：

① データの抽出が容易・柔軟に行なえる
② 文献の内部構造を反映している
③ データの流通と共有が容易である

本稿では，これらの条件を満たす方法として XML を用いて電子データの記述を行う．その理由としては，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．

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，既に欧米ではコーパスの作成に XML が用いられ始めており，今後も普及が予想されるという（齊藤俊雄ほか 2005:37）．

本稿の筆者は既にいくつかの朝鮮語史資料を 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．また須賀井義教(2009)では，15 世紀の朝鮮語文献を XML によりデータ記述する際の問題点として，構造化の基本となる単位をどのようにとらえるか，傍点や入力できない文字をどのように取り扱うかといった点を挙げた．本稿はこれらの成果物，検討結果に立ったものである．

2. データ記述のための枠組み

文献資料を構造化しデータとして記述する際，構造化の基準となる単位をどのように設定するかという問題がある．須賀井義教(2009)では，文献の物理的構造（張次や張の表裏，行数など）を構造化の基準とするか，あるいは「文」「段落」といった，文献内容の論理構造に着目して構造化を行うかという点について検討を行った．前者の例として，筆者が作成・公開している XML 文書が挙げられる．また後者の例としては，21 世紀世宗計画で配布している TEI を用いたデータ記述や，「太陽コーパス」などがある．

検討の結果，検索やデータの抽出などといった操作を行う上では，張や行といった物理的構造よりも，内容の論理構造を基本として構造化を行うのが便利であると考えられる．本稿では「太陽コーパス」にならい，「文」を基本として，張の切れ目などを適宜挿入していく方法をとる．

XML を用いて電子データを記述する際，本稿では以下の要素・属性を設定する（属性のうち，［ ］でくくったものは随意的なものである）：

表 1：本稿で設定するタグの一覧

要素名	属性	備考
book（文献）	title（書名），［vol（巻次）］	ルート要素
info（書誌情報）		book の下位要素
year（刊年）		info の下位要素
author（著者）		
copy（影印情報）		
text（本文）		book の下位要素
s（文）		text の下位要素
note（注釈）		s の下位要素
quote（会話）	［sp（話者）］，［h（聴者）］	
pb（張次）	no（番号）	
gap（欠落）	extent（欠落の分量），unit（単位：字・行など）	

上の表に見るように，欄外や本文への書き込み，墨書による校正といった情報は現時点で含んでいない．こうした付加的な情報については今後検討し，必要なものはさらに盛り込んでいきたい．また，注釈の中に現われた注釈や，経典の本文とそれに対する注釈といった，テクストのレベルの違いを反映する要素は設定しなかった．

表 1 に示したタグの一覧は，文献の内容を大まかに構造化したものであり，この枠組みを基本として，利用者によって必要な情報を要素としてさらに設定し，再利用することが可能である．

3. XML によるデータ記述の実際

ここでは『釈譜詳節』巻六を資料として，データ記述の実例について見てみる．まずは全体の概略を示す：

例：『釈譜詳節』巻六のデータ例

```
 1:  <?xml version="1.0" encoding="UTF-8"?>
 2:  <book title="釋譜詳節" vol="6">
 3:      <info>
 4:          <author>首陽大君</author>
 5:          <year>1447</year>
 6:          <copy>세종대왕기념사업회(1991), "역주 석보상절 제 6·9·11"</copy>
 7:      </info>
 8:      <text>
 9:          <s>
10:              <pb no='01a' />
11:              釋譜詳節第六
12:          </s>
13:          <s>
14:              世尊이 象頭山애 가샤 龍과 鬼神과 위ᄒᆞ야 說法ᄒᆞ더시다.
15:              <note>
16:                  龍鬼 위ᄒᆞ야 說法ᄒᆞ샤미 부텻 나히 셜흔 둘히러시니 穆王 여슷찻 히 乙酉ㅣ라.
17:              </note>
18:          </s>
19:          <s>
20:              부톄 目連이ᄃᆞ려 니ᄅᆞ샤ᄃᆡ
21:              <quote sp='부텨' h='目連'>
22:                  네 迦毗羅國에 가아 아바님긔와 아ᄌᆞ마님긔와
23:                  <note>
24:                      아ᄌᆞ마니믄 大愛道ᄅᆞᆯ 니르시니 大愛道ㅣ 摩耶夫人
25:                      <pb no='01b' />
26:                      ㅅ 兄니미시니 양ᄌᆡ 摩耶夫人만 몯ᄒᆞ실ᄊᆡ 버근 夫人이 ᄃᆞ외시니라.
27:                  </note>
28:                  아자바님내ᅀᅴ 다 安否ᄒᆞᅀᆞᆸ고 [中略] 涅槃 得호ᄆᆞᆯ 나 ᄀᆞᆮ게 ᄒᆞ리라.
29:              </quote>
30:          </s>
31:          [中略]
32:      </text>
33:  </book>
```

上の例では，「book」要素を最上位のルート要素として，以下「info」「text」の要素が続く（3行目，8行目）．「book」要素の属性には，データを抽出する際のラベルとして使用する書名，巻次を記述した（2行目）．巻次が不必要の場合には「vol」要素は記述しない．

「info」要素には著者，刊年などの書誌に関する情報を含める（4～6行目）．

「text」要素以下が文献の本文である．本文では終結語尾の出現を基準として「文」とみなし，「s」要素でマークした．ただし，ここでは会話の「quote」要素，注釈の「note」要素内に現われる「文」については「s」でマークしていない．「quote」要素について，21行目にあるとおり，話者と聴者が明示されている場合，あるいは文脈から明確に分かる場合には，それぞれ属性として記述した．単純な会話だけでなく，心に思ったこと，独話などについてもマーク付けを行った．以下の例では「太子」が心に感じた内容について，「quote」タグを付してある：

例：「quote」要素の例（『釈譜詳節』巻六 24a-24b）

```
<s>須達이 다시곰 請ᄒᆞᆫ대 太子ㅣ 앗겨 ᄆᆞᅀᆞ매 너교ᄃᆡ <quote sp='太子'>비들 만히 니르면
<pb no='24b' />몯 삵가?</quote> ᄒᆞ야 닐오ᄃᆡ [以下略]</s>
```

また，張の切れ目で「pb」要素を挿入した．これは「no」属性のみを持つ空要素である．「no」属性には張番号と，a・b によって張の表裏を記述した．

文字や行，張の欠落などを示す「gap」要素については，以下の例の通りである：

例：「gap」要素の記述例

```
草木이며 <gap extent='1' unit='文字'>■</gap>디며
                                        (『釈譜詳節』巻六 43b)
이 法師品은 經 <pb no='26a' /><gap extent='4' unit='行'>■■■</gap></note></s>
                                        (『釈譜詳節』巻十九 26a)
```

上の例のうち，「草木이며 …」の例は 1 文字が欠落している，あるいは不鮮明で判読できない場合であり，「이 法師品은 …」の例は行の単位で欠落している場合である．欠落を本文中に示すため，便宜上「■」を要素の内容として入れてある．

本稿ではこのようにして記述した電子データを，さまざまな形に加工して利用する方法をいくつか提示した．例えば XSLT（XSL 変換）を用いてウェブブラウザでの閲覧に適した形式に変換したり（図 1），注釈部分を折りたたんで表示するような形式に変換を行った（図 2）．

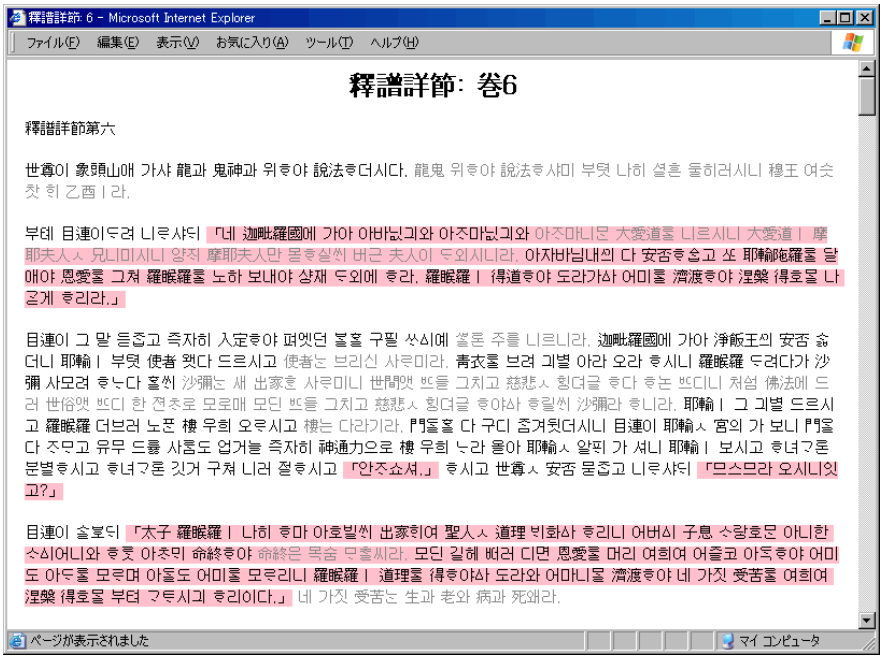

図 1：着色を施した表示形式

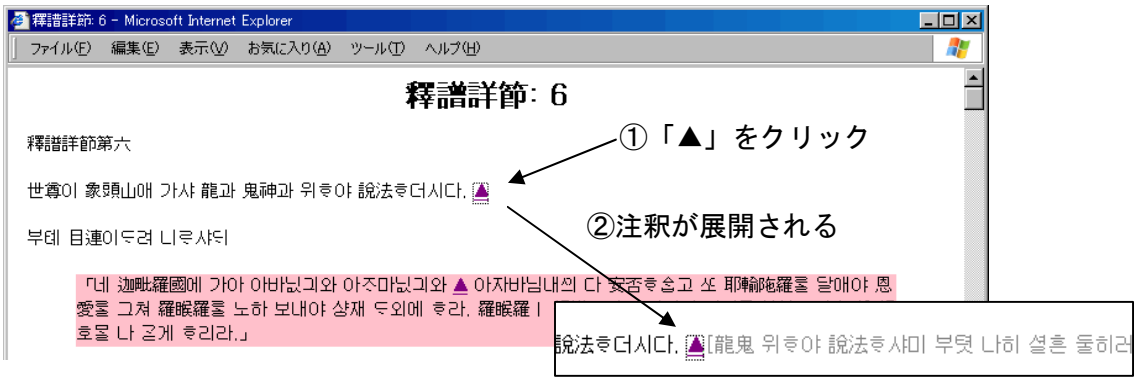

図 2：注釈の折りたたみ表示と展開

ウェブブラウザを表示に用いることで，このように様々な表現が可能となる．こうした形式への変換を容易に行うことができるのが，XML の特徴である．また，タグを全て取り去った単純なテキストの形式にすることももちろん可能であるが，ここでは省略する．

会話文に付与した「quote」要素のタグを利用して，出現箇所も含めた会話文のリストを作成することもできる（図 3）．

釋譜詳節: 6 会話リスト

書名:巻次	位置	話し手	聞き手	前	内容	後
釋譜詳節: 6	01a-01b	부텨	目連	… 目連이ᄃᆞ려 니ᄅᆞ샤ᄃᆡ	「네 迦毗羅國에 가아 아바ᄂᆡᆷ긔와 아ᄌᆞ … ᄆᆞᆯ 나 곧게 ᄒᆞ리라. 」	▲
釋譜詳節: 6	03a	耶輸	目連	… 구쳐 니러 절ᄒᆞ시고	「안ᄌᆞ쇼셔. 」	ᄒᆞ시고 世尊ㅅ
釋譜詳節: 6	03a	耶輸	目連	… 좀 묻ᄌᆞᆸ고 니ᄅᆞ샤ᄃᆡ	「므스므라 오시니잇고? 」	▲
釋譜詳節: 6	03a-04a	目連	耶輸	目連이 ᄉᆞᄅᆞᄃᆡ	「太子 羅睺羅ㅣ 나히 ᄒᆞ마 아ᄒᆞᄫᆞᆯ씬 … 그ᄐᆞ시고 ᄒᆞ리이다. 」	
釋譜詳節: 6	04a-06a	耶輸	目連	耶輸ㅣ 니ᄅᆞ샤ᄃᆡ	「如來 太子ㅅ 時節에 나ᄅᆞᆯ 겨집 사ᄆᆞ … 므슴 慈悲 겨시거뇨? 」	ᄒᆞ고 目連이ᄃᆞ려
釋譜詳節: 6	04b			… 보내샤 盟誓ᄒᆞ샤ᄃᆡ	「道理 일워사 도라오리라. 」	ᄒᆞ시고 鹿皮옷
釋譜詳節: 6	05b				「太子ㅣ 道理 일우샤 ᄌᆞ개 慈悲호라. 」	ᄒᆞ시ᄂᆞ니 慈悲ᄂᆞ
釋譜詳節: 6	06a	耶輸	目連	… 目連이ᄃᆞ려 니ᄅᆞ샤ᄃᆡ	「도라가 世尊ᄭᅴ 내 ᄠᅳ들 펴아 ᄉᆞᄅᆞᄫᆞᆸ쇼 … 」	▲
釋譜詳節: 6	06b	淨飯王	大愛道	… 道ᄅᆞᆯ 블러 니ᄅᆞ샤ᄃᆡ	「耶輸는 겨지비라 法을 모ᄅᆞᄆᆞᆫ씬 ᄆᆞᆯ굳 … 아 아라ᄃᆞᆯ게 니르라. 」	▲
釋譜詳節: 6	07a-07b	耶輸	大愛道	大愛道ᄭᅴ ᄉᆞᄅᆞᄫᆞᆸ샤ᄃᆡ	「내 지ᄇᆡ 이싱 젹기 뎌ᄃᆞᆯ ᄂᆞᄅᆞᆷ 天이 … 시ᄂᆞ니 엇더ᄒᆞ니잇고? 」	▲

図 3：会話リスト（一部）

図 3 に見られるように，会話の直前・直後についても部分的に抽出し，文脈つき索引のような形でリストを作成することができる．こうした会話文のリストを話者・聴者の情報も合わせて作成することで，身分の違いによる待遇表現使用の様相や，発話の形式などを確認することができよう．

本稿では試みに，用言についてのみ品詞情報の付与を部分的に行った．品詞情報を記述しておくことで，その情報を基にして，データの抽出をより柔軟に行うことができる．例えば特定の語尾を含む語節を抽出したり，資料に現れる形態の一覧を作成したりすることなどが可能となり，計量的な研究にも役立つものと考えられる．

用言に対し「v」要素を指定し，語幹（「st」），先語末語尾（「pfe」），語末語尾（「end」）を属性として記述した．例えば以下の例のとおりである：

例：「v」要素のマーク付け例（『釈譜詳節』巻六冒頭部分）

世尊ᄋᆡ 象頭山애 <v st='가다' pfe='시' end='아'>가샤</v> 龍과 鬼神과 <v st='위ᄒᆞ다' end='아'>위ᄒᆞ야</v> <v st='說法ᄒᆞ다' pfe='더/시' end='다'>說法ᄒᆞ더시다</v>.

媒介母音「-ᄋᆞ/으-」については記述しない．また語末語尾「-아/어」のように，母音調和に伴う異形態を持つ場合には，陽母音につく語尾を採択した．複数の先語末語尾が含まれる場合には「/」（スラッシュ）で区切ることとした．

本稿では試みに用言を含む語節についてのみマーク付けを行ったが，今後品詞情報のマーク付けを本格的に行う場合の問題点として，品詞分類の枠組み，異形態の扱いなどを挙げることができる．

4. おわりに

本稿では，XML による朝鮮語史資料の電子データ化について，15 世紀の文献をいくつか例として実際のデータ記述とその枠組みを紹介した．XML でデータ記述を行うことで，データをさまざまな形式へ変換し，情報を抽出できることが利点である．本稿で紹介した事例はごく一部であって，今後大量のデータを作成，蓄積していくことで，さまざまな研究に役立てることができよう．

本稿で扱うことのできなかった問題点として，「文」の認定をどのように行うか，「傍点」をどのように扱うかといった点が挙げられる．

さらに今後の課題として，データを利用するためのツール作成，品詞情報タグの付与に関する具体的な作業方法の検討などがある．また，言語資料として用いるためには，今後より多くの文献のデータ化が必要であり，データ記述の枠組みをさらに検討し，補充していきたいと考える．

参考論著

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）．

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．

齊藤俊雄・中村純作・赤野一郎(2005)『英語コーパス言語学――基礎と実践――』（改訂新版），東京：研究社．

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

須賀井義教(2009)「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」，『語学教育部ジャーナル』第 5 号，大阪：近畿大学語学教育部，pp.75-89.

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.
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（A03）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).
강범모(2003) "언어, 컴퓨터, 코퍼스 언어학" , 서울: 고려대학교 출판부.
서상규・한영균(1999) "국어정보학 입문", 서울: 태학사.
安秉禧(1982;1992) '국어사 자료의 書名과 卷冊에 대하여', "國語史 資料 硏究", 서울: 文學과 知性社.
安秉禧・李珖鎬(1990) "中世國語文法論", 서울: 學硏社.
李浩權(2001) "석보상절의 서지와 언어", 서울: 태학사.
조의성(2002) "月印釋譜(卷一) 語彙索引", 서울: 도서출판 박이정.
홍윤표(2005) '국어사 연구를 위한 전자자료 구축의 현황과 과제', "국어사 연구 어디까지 와 있는가"(국어사 학술 발표대회 발표 요지), 서울: 연세대학교 국학연구원.
Lehmann, H. M., C. Keller and B. Ruff(2006) 'ZEN Corpus 1.0', Facchinetti & Rissanen (eds.) *Corpus-based Studies of Diachronic English*, Bern: Peter Lang AG.
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).

参考ウェブサイト

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/