Errata and Corrections to Abstracts Proceedings, Digital Humanities 2014

Dear Lecturers,

Due to technical issues related to the large amount of contributions in addition to manual processing constraints, a few mistakes could only be identified after this book was sent out to print. The following pages are intended as a means to correct the most apparent misprints.

If it is needed, you may also refer to the full PDF version, available for free on www.dh2014.org, in which all known issues have been thoroughly resolved.

Should you find any additional mistakes that have not been taken care of, please feel free to inform us.

We sincerely apologize to all the people whose abstracts were affected.
Errata and Corrections to Abstracts Proceedings, Digital Humanities 2014

Table of Contents
Page x, section Papers, line 10: Andrew, Liam, see "Liam, Andrew" (page xiii)
Tomasek, Kathryn, see "Baumari, Syd" (page x)

Page xiii, lines 3-4: Koolen, Carina et al., Beautiful lips and porcelain cheeks appears twice. Second line should be Jackson, Cornell Alexander, Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland.

Page xiii, line 15: "Liam, Andrew" should be "Andrew, Liam and Gonzalez, Desi"

Missing Entries:
- Jackson, Cornell Alexander, Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland; page 237
- Kawase, Akihiro, Ichimura, Taro, Ogiso, Toshinobu, Problems in Encoding Documents of Early Modern Japanese: full paper missing, enclosed here.

Papers
Page 73: Paper title should read Extraction instead of Discovery.
Page 94: Formulas are missing in section 2.2. Actual text should read:

[...] After analyzing his choices, I produced the following metric \( w(k, n) \) to give a weight (to each nearby word \( n \) for the key word \( k \):

\[
\begin{align*}
\text{Let } w(k, n) &= \begin{cases} 
4^{v(k, n)} & \text{if } v(k, n) \geq 0 \\
-4^{v(k, n)} & \text{if } v(k, n) < 0 
\end{cases} \\
v(k, n) &= c_r + p(k, n) + c_t + f(n) + s(k, n) \\
p(k, n) &= \text{the number of punctuation marks between } k \text{ and } n; \text{ (contiguous punctuation marks count as 1)} \\
f(n) &= \begin{cases} 
1 & \text{if } n \text{ is a content word} \\
0 & \text{if } n \text{ is a function word}
\end{cases} \\
s(k, n) &= 10 \times 0.93^{(a_s + d_s + a_k + d_k + c_p + c_r)} \end{align*}
\]

Let \( a_k \) be the nearest common ancestor of \( k \) and \( n \) in the phrase structure tree and \( a_n \) be the nearest common ancestor of \( k \) and \( n \) in the dependency tree. Then \( d_{pk} \) is the distance between \( a_k \) and \( k \) in the phrase structure tree, \( d_{nk} \) is the distance between \( a_n \) and \( n \) in the phrase structure tree, \( d_{ak} \) is the distance between \( a_k \) and \( k \) in the dependency tree, and \( d_{nk} \) is the distance between \( a_n \) and \( n \) in the dependency tree.

Each possible context is evaluated as the sum of \( w(k, n) \) for each \( n \) in the possible context; the context with the highest value is chosen. If multiple possible contexts have identical values, any can be chosen; I picked the one that had the most context before the key word.

The constants were optimized to the following values using a Monte Carlo particle filter on the training data:

\[
c_f = 1.5; c_p = -3.37; c_{pk} = 0.175; c_{pm} = 0.2; c_{dk} = 1; c_{dn} = 1.4.
\]
Problems in Encoding Documents of Early Modern Japanese

Kawase, Akihiro
a_kawase@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

Ichimura, Taro
tichimura@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

Ogiso, Toshinobu	
togiso@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

1. Introduction

As part of its KOTONOHA (meaning ‘words of the language’ in classical Japanese) Project, the National Institute of Japanese Language and Linguistics (NINJAL) is conducting a morphological analysis of Japanese classics. Both the selection and digitization of classic documents are required in order to execute the analysis correctly. Digitization and morphological analysis have been done thus far on the literature of several ages and styles (Ogiso et al., 2012; Ichimura et al., 2012, 2013), and various text corpora have been published. These text corpora are marked up with NINJAL’s original Document Type Definition (DTD). However, some elements are used in common on all corpora but basically are not unified nor standardized. Under this circumstance, it enables structural analysis and string extraction from a single corpus but causes problems with structural comparison and numerical analyses between several corpora. Thus, it is necessary to design and mark up a unified definition from a higher level in order to conduct analyses concurrently.

In the previous study, we examined the possibility of converting classic documents with TEI-compliant XML (Kawase et al., 2013), in particular, by designing a tag-set and strictly structuring an old wood-block printed book from Sharebon’s “Keisei-kai Futasuji-no-michi” (published in 1798, e.g. Fig. 1) as a model case, and sorted out this problem. This work aims to provide further insights into those problems and propose a solution.

Fig. 1: Excerpt image of Sharebon’s “Keisei-kai Futasuji-michi” (owned by NINJAL)

2. Significance of Encoding Classic Documents

The reason for choosing Sharebon as a subject is because it is important primary linguistic data from the early modern period (Keene, 1995) and it has the following three features: (1) published in the broad age from the 18th to the first half of the 19th century; (2) uses colloquial words and expressions among conversations between characters; (3) abundantly describes things in Edo (former name of Tokyo) language and Kamigata (Kyoto-Osaka region) language.

Therefore, describing Sharebon in a machine-readable manner has profound significance not only for bringing numerical analysis of unexplained colloquial expressions of the modern period into reality but also for facilitating humanities research in language history and descriptive bibliography. Furthermore, since many texts that have a similar structure to Sharebon were published over the same period of time, this study will offer a common format for archiving coeval literature.

3. Issues in Encoding Sharebon

In order to analyze the manuscript from a corpus linguistic viewpoint, not only outward mark-ups but also internal mark-ups of both document structure and linguistic structure are required. In general, Sharebon is composed of a combination of three parts: the front matter; the narrative body that contains colloquial expressions and descriptive texts; and the back matter. For instance, in the case of “Keisei-kai Futasuji-michi”, the narrative body is made up of two chapters, ‘Natsu-no-toko’ (Summer Alcove) and ‘Fuyu-no-toko’ (Winter Alcove). Since the composition of this whole document is well accorded with the composition of an orthodox Western manuscript, we may mark up most of the elements in reference to TEI P5: Guidelines (Burnard and Baumann, 2007). Table 1 shows the list of elements with an explanation of their roles used to mark up the document.

The problems incurred in the process of encoding Sharebon can be broadly classified into two matters: (A) text formatting on a #PCDATA (character data) level; and (B) structuring ruby annotations. We will discuss these two matters in more detail in the following sections.

<table>
<thead>
<tr>
<th>Element</th>
<th>Role</th>
<th>Element</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;text&gt;</td>
<td>whole text</td>
<td>&lt;root&gt;</td>
<td>sentence unit</td>
</tr>
<tr>
<td>&lt;front&gt;</td>
<td>front matter</td>
<td>&lt;li&gt;</td>
<td>verse line</td>
</tr>
<tr>
<td>&lt;body&gt;</td>
<td>body part</td>
<td>&lt;ref&gt;</td>
<td>referencing string</td>
</tr>
<tr>
<td>&lt;back&gt;</td>
<td>back matter</td>
<td>&lt;ruby&gt;</td>
<td>ruby annotation*</td>
</tr>
<tr>
<td>&lt;div&gt;</td>
<td>divisions</td>
<td>&lt;ext&gt;</td>
<td>extended character</td>
</tr>
<tr>
<td>&lt;head&gt;</td>
<td>title</td>
<td>&lt;pb&gt;</td>
<td>page break</td>
</tr>
<tr>
<td>&lt;p&gt;</td>
<td>paragraph</td>
<td>&lt;cb&gt;</td>
<td>column break</td>
</tr>
<tr>
<td>&lt;q&gt;</td>
<td>quote</td>
<td>&lt;lb&gt;</td>
<td>line break</td>
</tr>
<tr>
<td>&lt;st&gt;</td>
<td>interlinear notes</td>
<td>#PCDATA</td>
<td>character data</td>
</tr>
</tbody>
</table>

Fig. 2: List of elements with an explanation of their roles to mark up Sharebon

3.1 Text Formatting on #PCDATA

To design and assure a high-quality corpus in terms of linguistic resources, it is necessary to index information about lexical morphemes (e.g. word class, inflected forms, pronunciations) at the body-text level. However, since there are three problems in the sentences of early modern Japanese, it is difficult to conduct morphological analysis properly: (A-1) voice markings called dakuten are missing where they should be; (A-2) corresponding phonetic characters called hiragana and katakana are not unified; (A-3) iteration symbols called otorii are unmodified.

For example, dakuten makes a phonetic shift from ka, ki, ku, ke, ko (written in hiragana letters as か, き, く, け, こ) to ga, gi, ge, go (rendered as が, ぎ, げ, ご). In vertical writing, ruby is typically printed on the right-hand side. However, since the three problems exist, it is difficult to mark up both outward

3.2 Structuring Ruby Annotations

Such annotations are the small-text characters rendered alongside the base text. In vertical writing, ruby is typically printed on the right-hand side. However, since the three problems exist, it is difficult to mark up both outward
information and the linguistic structure simultaneously: (B-1) ruby text carries extended characters (gaaji), damaged and missing characters, and some typographical errors and omissions; (B-2) words of such text and base text are not in one-to-one correspondence; (B-3) There can also be another ruby annotation on the left-hand side simultaneously.

4. Means for Solving the Problem

4.1 Text Formatting on #PCDATA

At NINJAL, for problems (A-1), (A-2), and (A-3), the character level is corrected using the original elements of \texttt{<vMark>}, \texttt{<kana>}, and \texttt{<odoriji>}, respectively (Ichimura et al., 2012, 2013). We accurately describe the text formatting in a uniform way by combining the TEI elements \texttt{<seg>} and \texttt{<choice>} (e.g. Fig. 2).

According to TEI P5: Guidelines (Burnard and Baumann, 2007), \texttt{<seg>} (arbitrary segment) represents any segmentation of text below the ‘chunk’ level, and \texttt{<choice>} groups a number of alternative encodings at the same point in the text. Distinctions in the corrections of (A-1), (A-2), and (A-3) can be shown by writing \texttt{@type}, holding selectable values from vMark, kana, and odoriji, to the attribute of element \texttt{seg}. The original text marked with TEI element \texttt{<orig>} (original form), along with the corrected text (the optimal version for morphological analysis) marked with TEI element \texttt{<reg>} (regularization) are written below the \texttt{<choice>} level.

This description policy preserves both the outward and linguistic structure, even if problems (A-1), (A-2), and (A-3) carry over extended characters or omissions in the original manuscript, and brings morphological analysis into reality as well.

4.2 Structuring Ruby Annotations

At NINJAL, regarding structuring ruby annotation, since priority is given to morphological analysis, words and characters are encoded using the original defined element \texttt{<ruby>} with an attribute \texttt{@rubyText} (e.g. Fig. 3a). To express this structure with TEI-compliant XML, we can simply substitute the element \texttt{<ruby>} with \texttt{<w>} (word) and the attribute \texttt{@rubyText} with \texttt{@ana} (e.g. Fig. 3b). However, the important ruby text information is described inside \texttt{@ana} as a value, we cannot solve problem (B-1) under this policy.

Depending on the technology, CSS, XHTHM, and HTML5 platforms might be considered as alternatives in order to structure ruby annotation (Benol, 2010) (e.g. Fig. 3c).

Here, each \texttt{<ruby>}, \texttt{<rt>}, and \texttt{<rp>} represent inline elements that contain base text with ruby annotation, the superscript which over comes the base text, and ruby parenthesis which are used to wrap around opening and closing parentheses \texttt{<rt>}, respectively. However, since the ruby text comes over the #PCDATA level, we cannot solve problem (B-2) under this policy. In addition, we have to solve the above problem and problem (B-3), securing the coexistence of ruby on the right-hand side and left-hand side, simultaneously.

4. Conclusion

In this study, developing a corpus for historical documents in a comprehensive and versatile way was considered the ultimate goal. We devised a tag-set based on the TEI-element, and structured Shaburori’s “Keisai-kai Futasujinchi” as a model case. We examined the problems of (A) text formatting on #PCDATA level and (B) structuring ruby annotations which were previously unsolved. For problem (A), a concrete policy to bring morphological analysis into reality by combing the element of \texttt{<seg>} and \texttt{<choice>}, was considered. For problem (B), major difficulties of structuring both outward information and the linguistic structure of Japanese documents at the same time were presented, and problems to solve were pointed out. Especially, ruby annotation is absolutely imperative for Japanese documents. This extra-textual addition is common and employed almost everywhere from historical documents to modern comics for highlighting the pronunciation or meaning of a word. Our future challenge is to examine a compromise that satisfies both structures simultaneously while working out the problems with the TEI Council.

5. Funding

This work was supported by the collaborative research project “Design of a Diachronic Corpus” and “Study of the History of the Japanese Language Using Statistics and Machine Learning” carried at the National Institute for Japanese Language and Linguistics.
References


