**ICHIMURA — Construction**

<7 figures>

**Construction of the Corpus of *Toraakira-bon Kyōgen***

Ichimura, T., Watanabe, Y., Kono, T. and Ogiso, T

As part of the Corpus of Historical Japanese, we are constructing a corpus of *kyōgen* books. In this corpus, we annotated both document structure information and morphological information for language study.

*Kyōgen* is a form of traditional Japanese comic theater. It developed alongside *Noh* in the Muromachi period (1337–1573). The *Okura Toraakira-bon* books are old script books, written by Okura Toraakira in 1642. As *kyōgen* uses an older form of the Japanese language, i.e., the language of the Muromachi period, these *kyōgen* books thus constitute important documents for researching colloquial language of the time.

Thus, we built the corpus as an original text in Otsuka (2006) that was the latest commentary of the *Okura Toraakira-bon* books and reflected the original well. They consist of 237 plays and 57 notes (approximately 270,000 words).

**Information on the Document Structure**

In *kyōgen* scripts, like in Western ones, we find lines to be spoken by the actors that were written in colloquial style, stage directions that were written in a literary style. To study the *kyōgen* scripts linguistically, it is necessary to distinguish the lines from the stage directions.

We studied tags and document type definitions for the *kyōgen* scripts (Table 1; see also Ichimura et al., 2012; Kobayashi et al., 2013) with reference to TEI P5 (TEI Consortium, 2007) and encoded them in XML files (Figure 1).

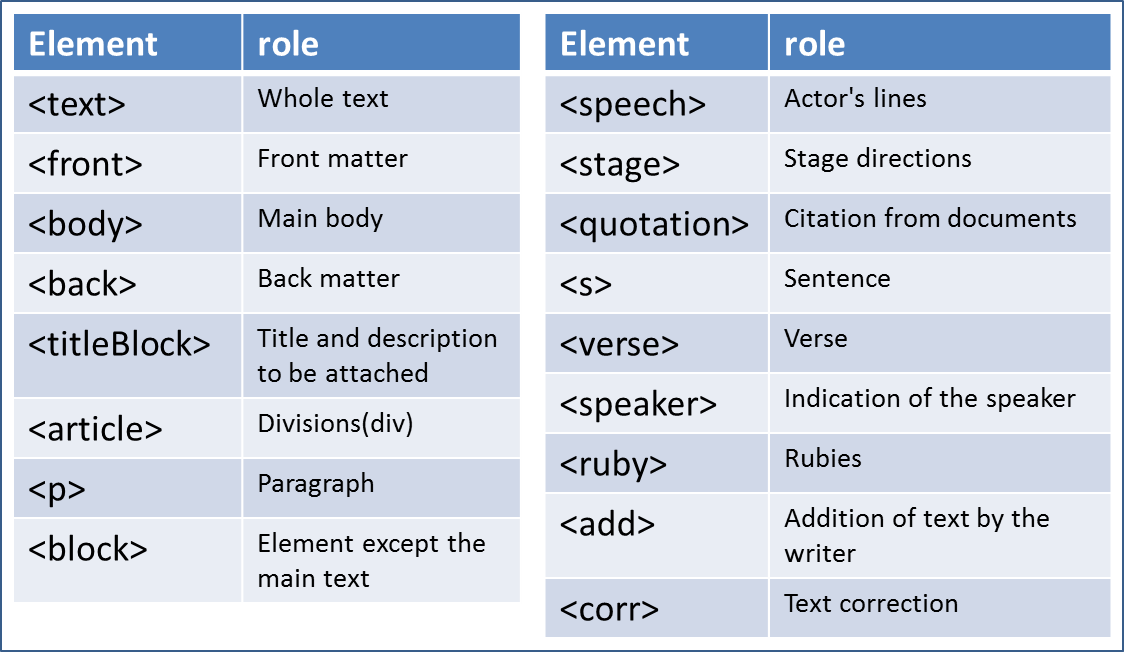


Table 1. Main tags.

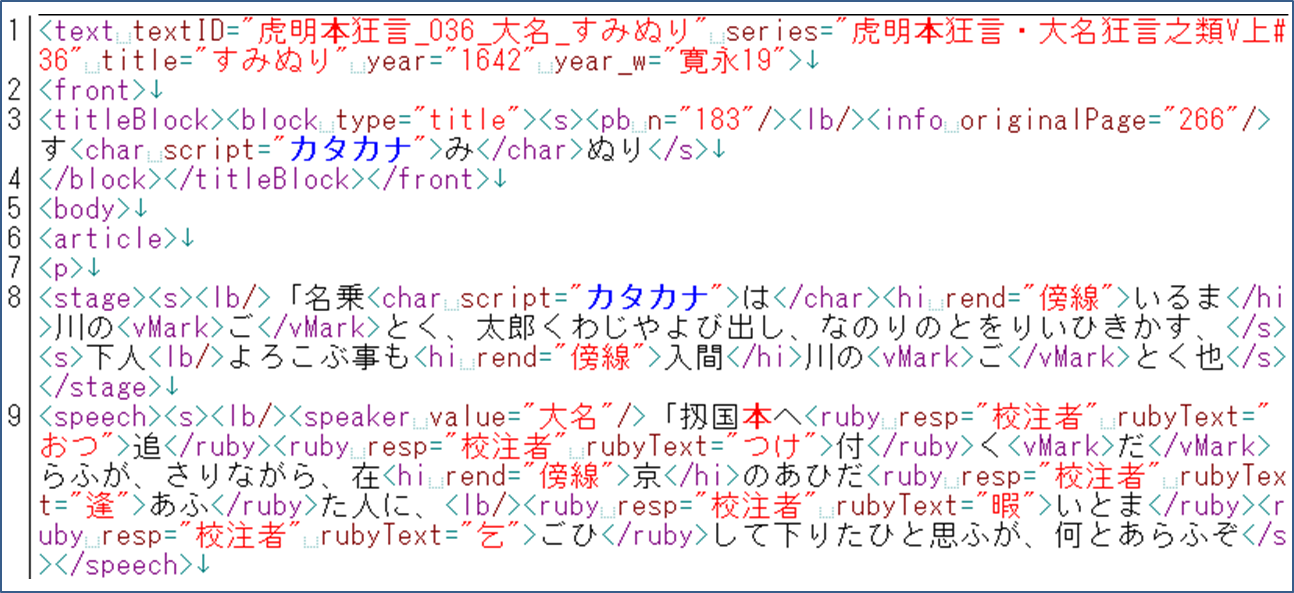


Figure 1. XML.

In the process, an issue about the notation arose. Because we used printed books that reflect the notation of the original manuscript, we were forced to increase the notation pattern in morphological analysis.

1. Three-letter words.

2. Whether to have a dot mark for a voiced sound or not.

3. Signs of the repetition.

To solve this problem, we revised and marked them according to the following policy.

1. We replaced the katakana with hiragana letters (Figure 2).

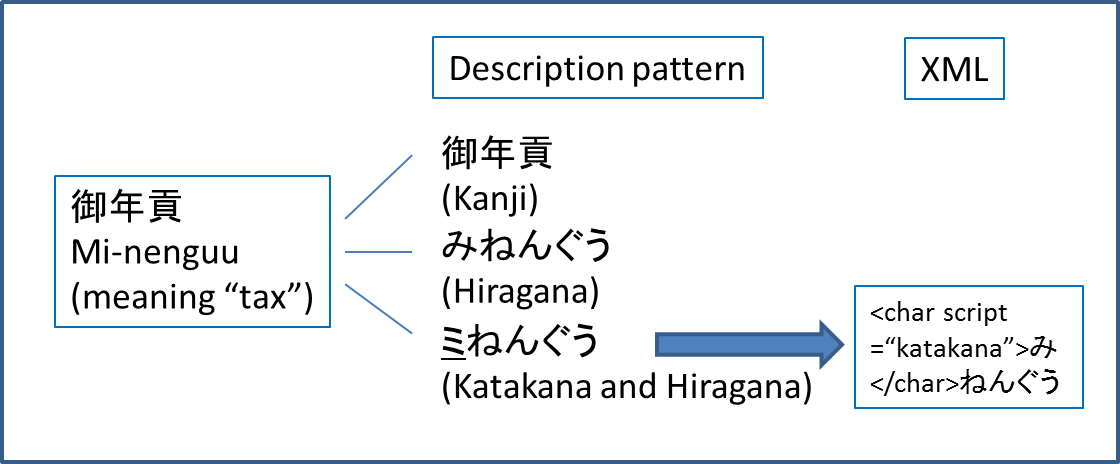


Figure 2. Revision of katakana.

2. We replaced a kana character containing a kana with a dot mark when it was clear that it was pronounced by a voiced consonant (Figure 3).

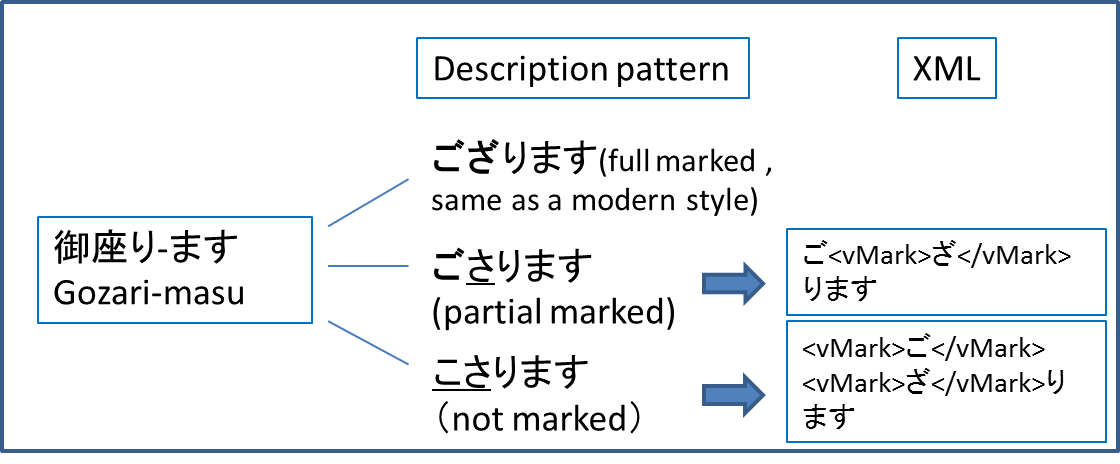


Figure 3. Revision of the kana scripts with the dot mark.

3. We replaced the signs for one character with the character applied to it. However, we did not replace signs for more than two characters, because their repeated range was wide and occasionally it spanned beyond a sentence (Figure 4).

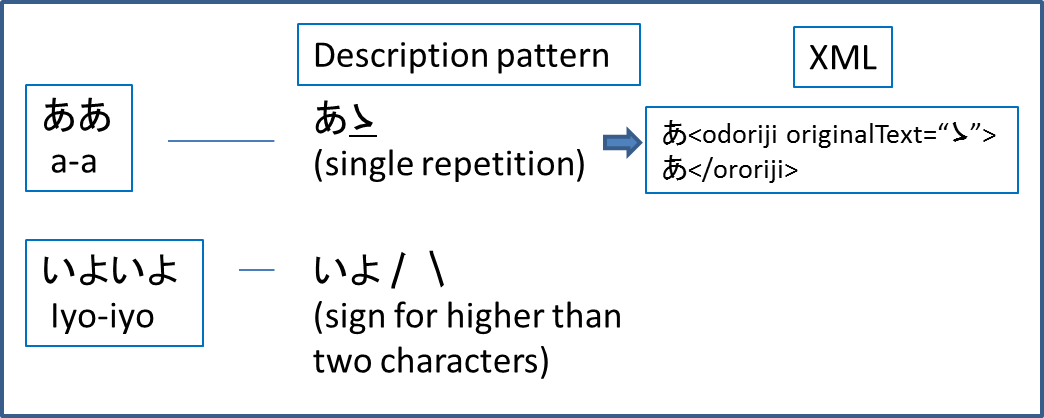


Figure 4. Revision of the repetition marks.

**Morphological Information**

In order to use it for language study, the corpus needs to be annotated to associate information to words.

In addition, significance can be obtained from morphological information. In written Japanese, there is no space between words like English, which was the case for early modern Japanese texts as well. In addition, Japanese has three kinds of letters or characters used in writing. Furthermore, in early modern times, spellings had not been fixed, which are fixed in contemporary Japanese (Figure 5).

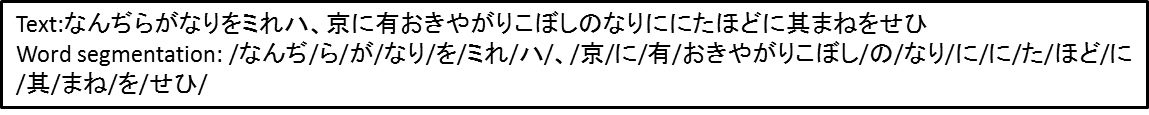


Figure 5. Word segmentation.

This makes it even more difficult to identify a desired word. To solve these problems, we used a morphological analysis system, and automatically annotated all words in the text with morphological information. *Kyōgen* texts were analyzed with a high degree of accuracy (an approximate 96% precision rate).

In this process, for each word, we gave the word information ‘lemma’, which unified notation and the conjugations of a word. Thus we came to be able to search the variation in various notations and conjugations of the desired word at a time.

**Utility Value of this Corpus**

The data are displayed on the search tool named *Chū-nagon* (Figure 6).

Using both morphological information and document structure information, we can search single words or series of words, including form variations, with the speakers and type of text.

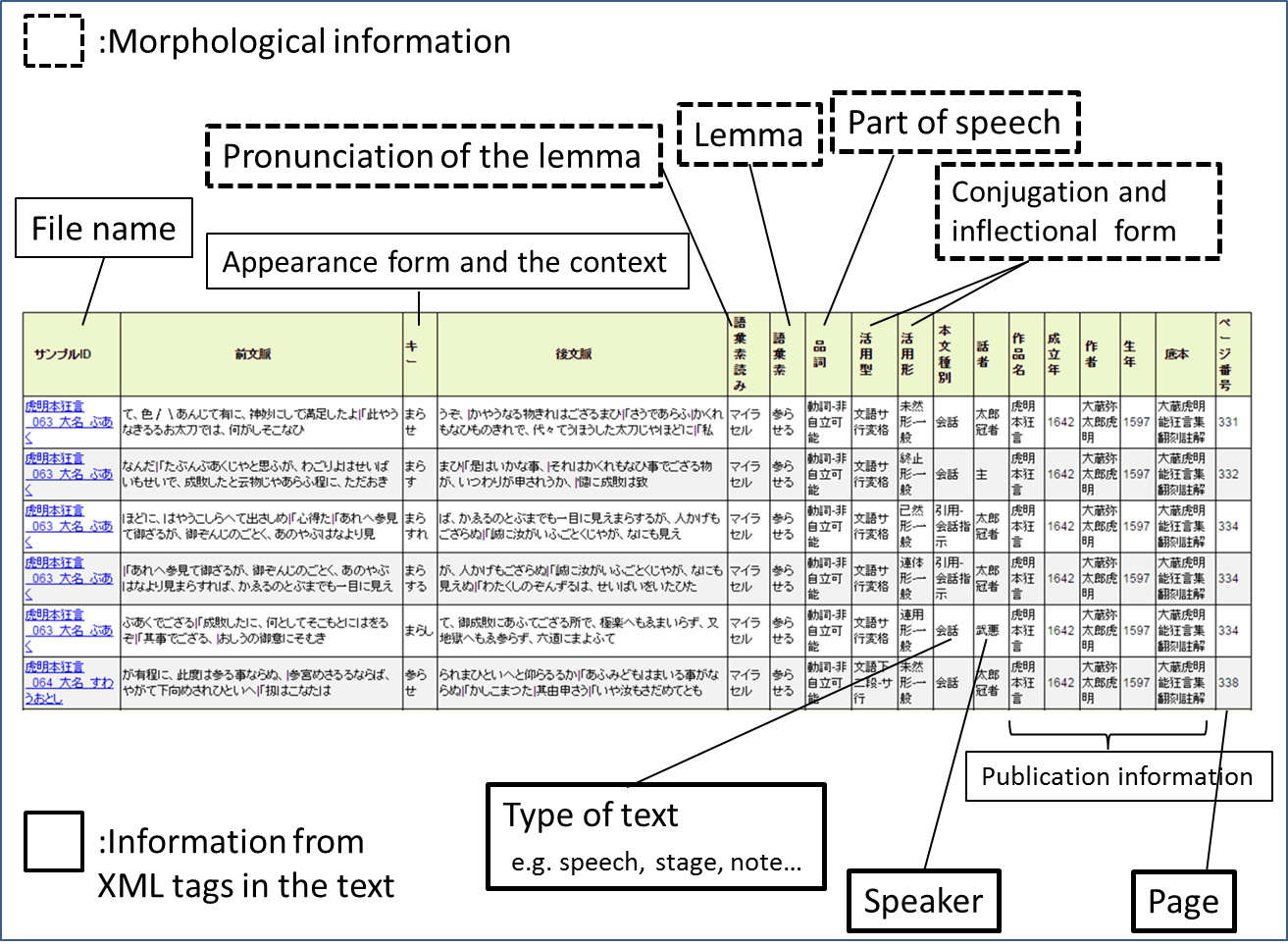


Figure 6. Search result on *Chū-nagon*.

We will be able to study the language of *kyōgen* in a more sophisticated manner.1 In addition, we can analyze what kind of expression was used by a certain character, or how the person was characterized in the drama.

NINJAL2 have already constructed the corpus of early middle Japanese (for the Heian period, or the years 794–1185) and early modern Japanese (for the Meiji period, 1868–1912). Now that we have constructed the high annotated corpus ofearly modern Japanese, we will be able to study the Japanese language from classical times to the present age diachronically using a computer.

**Note**

1. For example, when a user searches a series of words ‘おりやらします’, the examples and speakers are displayed as a list, and the person will notice that almost all the speakers of ‘おりやらします’ are women. Then the user will understand that this series of words is female language of the early modern Japanese.

2. NINJAL is an abbreviation for the National Institute for Japanese Language and Linguistics.

**References**

**Ichimura, T., Kawase, A. and Ogiso, T.** (2012). Structuring Colloquial Early Modern Japanese Text and Its Issues of Definition. *Proceedings of the 96th IPSJ SIG Computers and the Humanities, CH96*, pp. 1–8.

**Kobayashi, M. and Ichimura, T.** (2013). Structuring the Corpus of *Toraakira-bon Kyogen*. *Proceedings of the 3rd Workshop on Corpus Japanese Linguistics,* pp. 323–32.

**National Institute for Japanese Language and Linguistics.** (2009). Center of Corpus Development, NINJAL. http://www.ninjal.ac.jp/corpus\_center/.

**Ogiso, T., Ichimura, T. and Kouno, T.** (2013). Preliminary Study of Morphological Analysis of Early Modern Japanese. *Proceedings of the 4th Workshop on Corpus Japanese Linguistics,* pp. 145–50.

**Otsuka, M.** (2006). Okura toraakira noh kyogenshu: honkoku chūkai. 1. Seibundo shuppan, Osaka.

**TEI Consortium.** (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative. http://www.tei-c.org/Guidelines/P5/ (accessed 19 February 2015).