# **OGISO — Morphological**

# <1 image, 1 Word table>

# **Morphological Analysis of Japanese *Kyōgen* Text**

Ogiso, T., Kono, T. and Ichimura, T.

As part of the corpus of historical Japanese, we are constructing a corpus of *kyōgen* books. *Kyōgen* is a comic Japanese theatrical form that developed alongside *noh* in the Muromachi period (1337–1573). *Ōkura-ryū* (Ōkuraschool) is one major school of *kyōgen*, and the *Ōkura Toraakira bon* books are their oldest script books, written by Okura Toraakira in 1642; however, these *kyōgen* preserve an older form of the Japanese language than those of the other schools, reflecting the language of the Muromachi period, and as a result they constitute important documents for research into the colloquial language of the time. Therefore, we have begun to construct a corpus of *kyōgen* based on the text of the *Ōkura Toraakira bon* books, which consist of 237 plays (approximately 270,000 words).

This corpus requires textual annotation for morphological information, such as part-of-speech, lexeme, and pronunciation information. Initially, we used an existing morphological analysis system to automatically annotate morphological information for all words in the text. However, this system could not analyze the *kyōgen* texts with sufficient precision because of difference of the linguistic characteristics.

Therefore, we decided to develop a new electronic dictionary for *kyōgen* text based on UniDic, an existing electronic dictionary for morphological analysis of Japanese (http://sourceforge.jp/projects/unidic/). We used two approaches: expansion of the entries in the existing UniDic, and annotation of a new *kyōgen* corpus as training data for morphological analysis.

## UniDic for Kyōgen

Starting from the existing UniDic, we extended word entries to address the problems of lexical, morphological, and orthographic differences. UniDic is structured with layered entries in order to treat words flexibly depending on the purposes of researchers. Figure 1 exemplifies the structured word indexes of UniDic. The Lemma layer treats words at abstract lemmatized level, like the entries in a general dictionary. The Form layer distinguishes allomorphs and different conjugations. Specification of conjugations is held in this layer. Finally, the Orthography layer is prepared to distinguish orthographic variants.

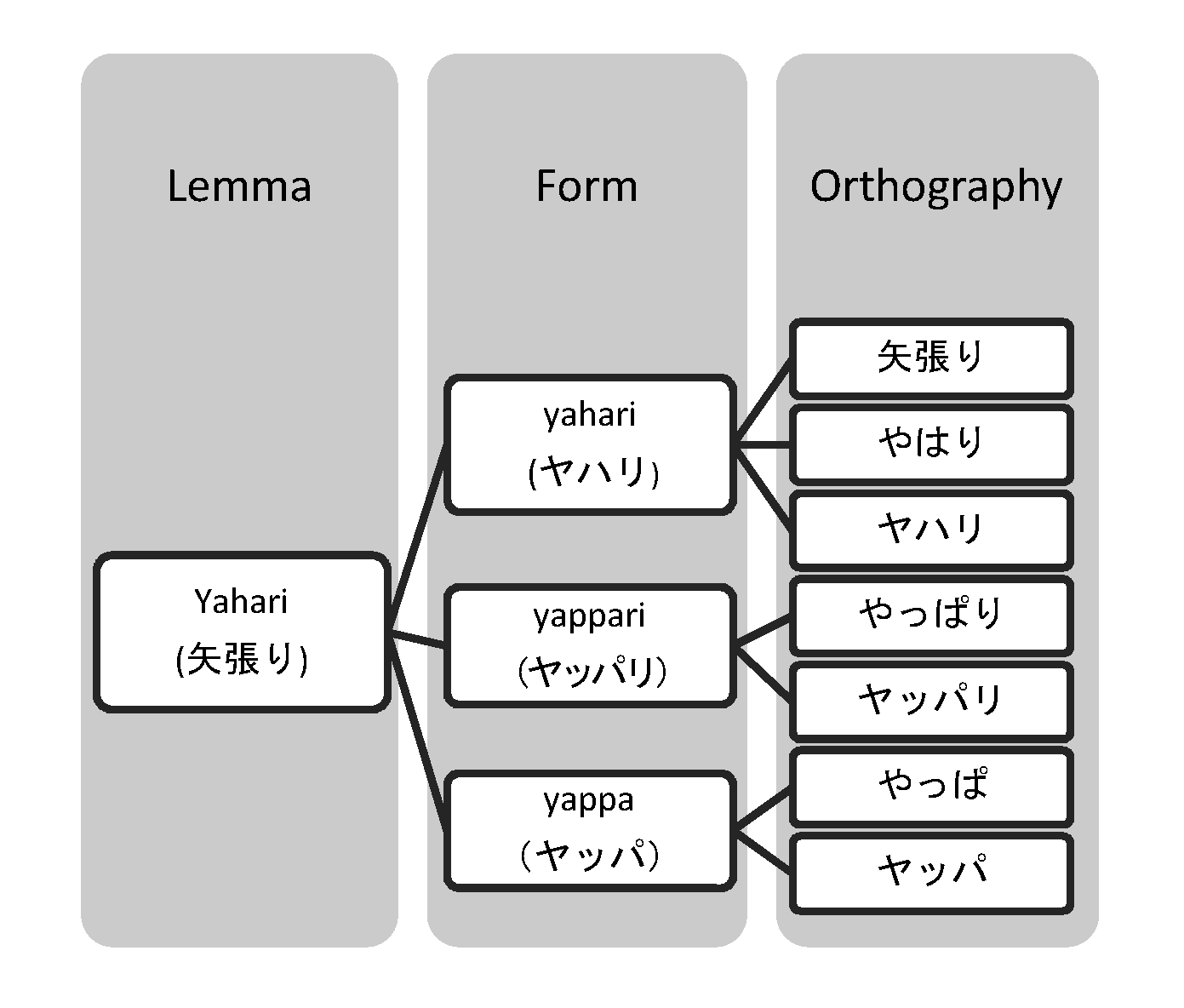


Figure 1. Hierarchical structure of UniDic.

We added new entries to each level in this structure, approximately 6,000 in all, to reflect these lexical, morphological, and orthographic differences.

To remedy the issues of morphological and syntactic differences between contemporary Japanese texts and *kyōgen* texts, we manually annotated a corpus of *kyōgen* containing 100,000 words in order to produce training and test corpora.

MeCab is a morphological analyzer based on the conditional random field (CRF) analytical method that achieves state-of-the-art performance in contemporary Japanese morphological analysis. MeCab can automatically learn feature weights for UniDic from an annotated corpus of *kyōgen* to build a morphological analyzer.

**Evaluation**

Using a dedicated dictionary (UniDic for *Kyōgen*), *kyōgen* texts can now be analyzed with a high degree of accuracy. We evaluated the performance of UniDic for *Kyōgen* using test data consisting of 10,000 words of randomly sampled sentences (10% of the annotated corpus). The evaluations were carried out on four levels. Level 1 covered the accuracy of word segmentation. Level 2 covered the accuracy of part-of-speech tagging for items correct at Level 1. Level 3 covered the accuracy of lemmatization for items correct at Levels 1 and 2. Finally, Level 4 covered the accuracy of distinction between allomorphs for items correct at all other levels. Table 1 shows the performance of UniDic for *Kyōgen* at each of the four levels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Level 1  Segmentation | Level 2  POS-tagging | Level 3  Lemmatization | Level 4  Allomorphs |
| F-measure | 0.9881 | 0.9629 | 0.9567 | 0.9536 |

Table 1. Accuracy of UniDic for *Kyōgen*.

**Conclusion**

The accuracy for lemmatization, which is the one most often used by linguists, is approximately 96%. This number is only slightly inferior in comparison with the accuracy of the morphological analysis dictionary for contemporary Japanese (approximately 98%). Using UniDic for *Kyōgen*, *kyōgen* texts can now be analyzed with a high degree of accuracy.

## References

**Den, Y., Nakamura, J., Ogiso, T., and Ogura, H.** (2008). A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation. *Proceedings of the 6th Language Resources and Evaluation Conference*, pp. 1019–24.

**Den, Y., Ogiso, T., Ogura, H., Yamada, A., Minematsu, N., Uchimoto, K. and Koiso, H.** (2007). The Development of an Electronic Dictionary for Morphological Analysis and Its Application to Japanese Corpus Linguistics (in Japanese). *Japanese Linguistics*, **22**: 101–23.

**Kudo, T.** (2012). MeCab: Yet Another Part-of-Speech and Morphological Analyzer. https://code.google.com/p/mecab/.

**Kudo, T., Yamamoto, K. and Matsumoto, Y.** (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–37.

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

**Ogiso, T., Komachi, M., Den, Y. and Matsumoto, Y.** (2008). UniDic for Early Middle Japanese: A Dictionary for Morphological Analysis of Classical Japanese. *Proceedings of the 8th Language Resources and Evaluation Conference*, pp. 911–15.