**KAWASE — Extracting**

<5 figures, includes a table appearing as a figure and some Japanese script in a box>

**The Characteristics of Personae Observed within *Toraakirabon Kyogen* Scripts: Extracting Conceptual Words Using Quantitative Analysis**

Kawase, A.

This study analyzes *Noh* farces (farces in the Japanese traditional performing arts), which have not been the subject of natural language processing research until now. The goal of this study is to quantitatively analyze the text of *Noh* farce scripts from the perspective of vocabulary to verify how the status and roles of the characters are defined. Specifically, this study compares the keywords (unique concepts) of the various statuses and roles by constructing a text corpus of *Noh* farce character dialogues categorized by status and role, and using co-occurrence networks of frequently occurring vocabulary and statistical indices in order to make the characteristics visible.

*Noh* farces are linguistic corpuses that have had an important position from medieval to early modern times (Keene, 1999). In particular, *Noh* farces have a very high value as a spoken-language corpus because of the following three points: (1) stories proceed in dialogue format, (2) the conversation portions clearly depict situations and statuses, and (3) characters are diverse and have different statuses and roles.

Therefore, organizing the text of *Noh* farce scripts into machine-readable formats and executing quantitative analysis not only contributes to the clarification of the actual language spoken in early modern times, but is also significant because it promotes humanities research, such as the history of the Japanese language and descriptive bibliography. This study provides a comprehensive analysis of character speech in *Noh* farces from a corpus linguistics perspective and deciphers the roles that the people play.

**Subject of Analysis**

The analysis in this study uses the text corpus in the *Toraakira Kyōgenshū* (*Toraakira Noh* Farce Collection, 1642), a collection digitized, organized, and morphologically analyzed by the National Institute of Japanese Language and Linguistics (NINJAL) (Ogiso et al., 2013). At NINJAL, digitization and morphological analysis have been done thus far on the literature of several ages and styles (e.g., Kawase et al., 2014a; 2014b). The analysis of character dialogue in this text corpus (e.g., Figure 1) was conducted using MeCab Ver.0.993 (Kudo, 2012) with the morphological analysis dictionary exclusive to *Noh* farces.

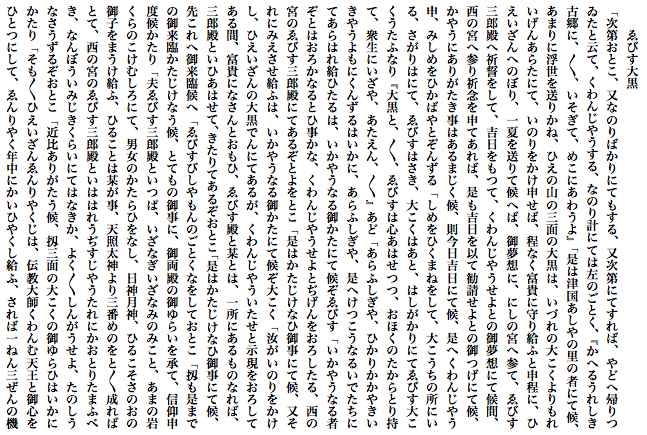


Figure 1. Excerpt script of *Ebisu Dai-koku* from *Toraakira Kyōgenshū*.

**Word Frequency Analysis**

Here, both TF-IDF (term frequency-inverse document frequency) and CIDF (comparative IDF) used in Murai and Tokosumi (2009) was applied to the *Noh* farce text corpus, and unique vocabulary was extracted for the status and role for the top 10 characters who spoke the highest number of times. *CIDF* is an index that defines the desired word *t* as a ratio of *idfa* (IDF value found based on the number of texts *da* where that word appeared in all text *Da*) and *idfg* (IDF value found based on the number of texts *dg* where that word appeared in all text *Dg* in the specific genre, which is the case in the statuses and roles of the characters):

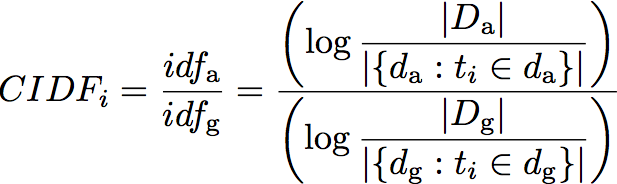
.

Table 1 shows the broader terms for the TF-IDF and CIDF value calculated for the nouns that appear in the dialogue between the *kaja* (servant) who appears most often in the *Toraakira Kyōgenshū* and the *shū* (master) who appears the next most often.

Tab1

Table 1. Extracted keyword lists for *kaja* and *shū* using TF-IDF and CIDF.

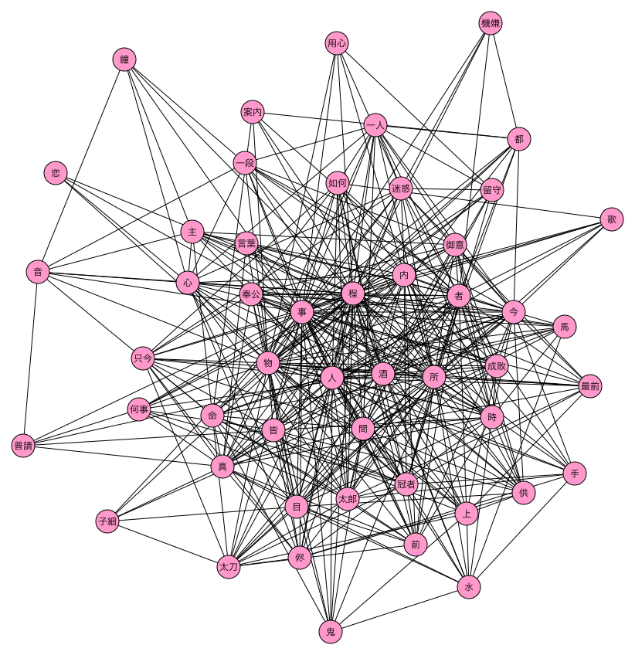


Figure 2. A keyword network (co-occurrence network) of *kaja*.

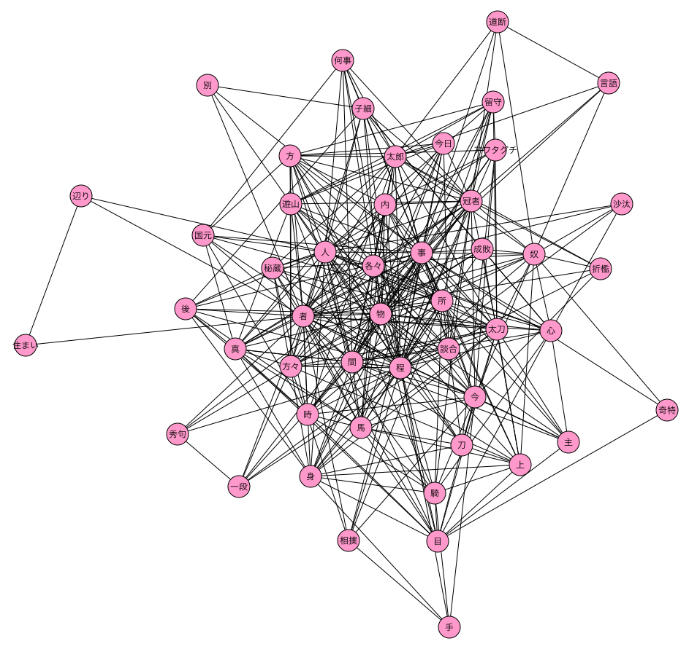
****

Figure 3. A keyword network (co-occurrence network) of *shū.*

**Network Centrality Analysis**

In order to explore the structure of the concepts for each character, we describe a co-occurrence network of the top 50 words that appeared in the dialogue (e.g., Figures 2 and 3), and applied network analysis techniques to numerically grasp the differences among the networks. More specifically, we used several networking algorithms based on the concept of centrality (e.g., degree, closeness, and eigen value). Table 2 shows the results of a summary of the top 10 values of network centrality for *kaja* and *shū*. As given above, effectively using the characteristic vocabulary extracted here gives the ability to define the status and role of the characters in the situation.

Tab2

Table 2. Values of network centrality for *kaja* and *shū*.

**Conclusions and Future Research**

Since *Noh farces* are for recreation and depict fictional worlds, and viewers understand people, events, and situations by integrating the information they obtain from the background knowledge and text (speech and behavior of the character) (van Dijik and Kintsch, 1983), the result of this study quantitatively clarified the characteristic vocabulary (references and concepts) that defines the statuses and roles of the characteristics in *Noh* farce script texts. However, *Noh* farces proceed in dialogue format, and the speech of the characters probably changes when the person listening to the speaker changes. A future topic is to extract the concepts for defining the statuses and roles of the characters more comprehensively and in more detail by analyzing the relationships between the speakers and the people listening, which will promote humanities research in the field of Japanese traditional performing arts.

**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**

**Kawase, A., Ichimura, T. and Ogiso, T.** (2014a). Problems in Encoding Documents of Early Modern Japanese. *Proceedings of the Digital Humanities 2014*, Lausanne, Switzerland, pp. 225–27.

**Kawase, A., Ichimura, T., and Ogiso, T.** (2014b). Quantitative Analysis of Dialogues in *Toraakira Kyōgenshū*. *Proceedings of the Annual Meeting of NLP 2014*, Baltimore, MD, pp. 662–65.

**Keene, D.** (1999). *A History of Japanese Literature*.Vol. 2. *World within Walls*. Columbia University Press, New York.

**Kudo, T.** (2012). MeCab: Yet Another Japanese Dependency Structure Analyzer. https://code.google.com/p/mecab/.

**Murai, H. and Tokosumi, A.** (2009). Towards a Quantitative Research Approach to Literary Review: Quantitative Analysis of Book Reviews. *Journal of Japan Society of Information and Knowledge,* **19**(2): 120–25.

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

**Van Dijik, T. A. and Kintsch, W.** (1983). *Strategies of Discourse Comprehension*. Academic Press, New York.