JADH2013 & DH-JAC2013 CONFERENCE

ABSTRACTS

19-21 September 2013

Ritsumeikan University, Kinugasa Campus,
Conference Room, Soshikan Hall, 1st Fl.

http://www.dh-jac.net/jadh2013/

Hosted by:
JADH 2013 Organizing Committee
Under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:
Art Research Center, Ritsumeikan University
Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University
Center for Evolving Humanities, Graduate School of Humanities and Sociology, University of Tokyo
International Institute for Digital Humanities

Co-sponsored by:
Japan Society of Information and Knowledge
The Mathematical Linguistic Society of Japan
IPSJ SIG Computers and the Humanities
Japan Association for East Asian Text Processing (JAET)
Japan Association for English Corpus Studies (JAECS)
Japan Association for the Contemporary and Applied Philosophy
Japan Art Documentation Society (JADS)

Lexical Modeling of *Yamabuki* (Japanese Kerria) in Classical Japanese Poetry

Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego)

This project is a lexical study of classical Japanese poetic vocabulary through network analysis based on graph theory. The analysis is based on co-occurrence patterns, defined as any two words appearing in a poem.

Many scholars of classical Japanese poetry have tried to explain the constructions of poetic vocabulary based on their intuition and experience. As scholars can only demonstrate constructions that they can consciously point out, those that they are unconscious of will never be demonstrated. When we develop a dictionary of poetic vocabulary using only our intuitive knowledge, the description will lack important lexical constructions. In order to conduct more exact and unbiased descriptions, it is necessary to use computer-assisted descriptions of poetic word constructions using co-occurrence weighting methods on corpora of classical Japanese poetry.

We developed the corpora of classical Japanese poetry based on the eight anthologies compiled under imperial order called the "*Hachidaishū*" which were established from ca. 905 to 1205. We also developed a method of co-occurrence weighting (Yamamoto, 2006) which calculates the weight of patterns of any two words appearing in a poem sentence similar to the *tf-idf* method (Sparck Jones, 1972; Robertson, 2004; Manning and Schutze, 1999). The CW allows us to examine the patterns of poetic word constructions through mathematical models.

As a result, when we draw a network model from co-occurrence patterns, we can in general observe a main hub node derived from a topic word. Additionally, we also encounter other hub nodes which do not indicate topic words nor entry items in a poetic dictionary. For instance, when we take *yamabuki* (Japanese kerria) as a topic word and draw its network model, we will observe *kahazu* (frog), *Ide* (place name, proper name), and *yahe* (eightfold or double ower) as hub nodes. The terms *yamabuki*, *kahazu*, and *Ide* are contained in some poetic dictionaries as entry items or collocations. The term *yahe* is, however, not seen in any poetic dictionaries even as a single term. We conclude that a term such as *yahe* can be shown as a hub node which takes an important role to connect a topic word with other peripheral words such as *kokonohe*, *nanahe*, *hitohe*, and plays a supporting role to form a poetic story in the poem even if it is not included in a dictionary.

The finding of this study is that the modeling developed here allows us to 1) discern not only patterns described by experts but also patterns yet undescribed, and 2) identify not only specific or tangible words but also abstract or conceptual words which have a tendency to be left out of dictionaries.

[Key Words: corpus linguistics, co-occurrence weight, visualization, Japanese literature, network modeling]

References

Manning, Christopher D. and Hinrich Schutze (1999) *Foundation of statistical natural language processing*, Cambridge, Massachusetts: The MIT press.

Robertson, Stephen (2004) "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, Vol. 60, pp. 503-520.

Sparck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, pp. 11{21.

Yamamoto, Hilofumi (2006) "Konpyūta niyoru utamakura no bunseki / A Computer Analysis of Place Names in Classical Japanese Poetry", in *Atti del Terzo Convegno di Linguistica e Didattica Della Lingua Giapponese, Roma 2005* : Associazione Italiana Didattica Lingua Giapponese (AIDLG), pp. 373-382.

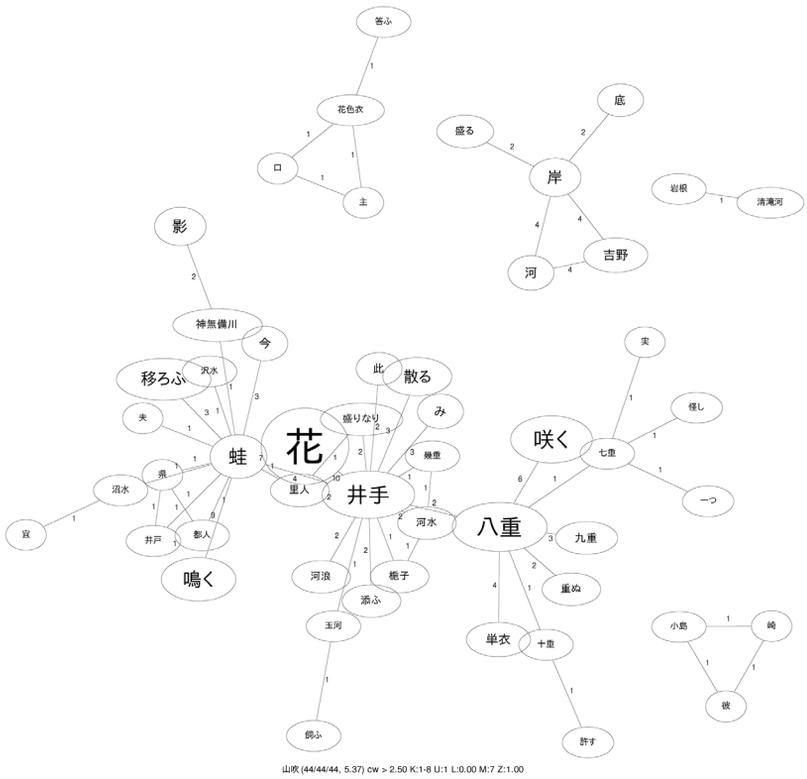

Figure 1: Graph model of *Yamabuki*: a core node, 山吹 *yamabuki*, is pruned. *kahazu* (蛙 , frog), *Ide* (井手 , place name, proper name), and *yahe* (八重 , eightfold or double ower) are observed as hub nodes.

A Diachronic and Synchronic Investigation into the Properties of Mid-Rank Words in Modern Japanese

Bor Hodošček (Tokyo Institute of Technology), and
Hilofumi Yamamoto (Tokyo Institute of Technology / University of California)

The present study focuses on the role of mid-rank words in modern Japanese. Mid-rank words are defined as words having an average TF-IDF (term frequencyinverse document frequency) score. Mid-rank words are often overlooked for words with high TF-IDF scores, which act as reliable topic markers. Words with low TF-IDF scores are in turn seen as functional words and often discarded from analysis. Mid-rank words are thus words that do not lean heavily towards the two extremes of topic and function, but include a mixture of both. As such, their exact grammatical function is elusive and still relatively unknown.

In order to determine the properties of mid-rank words, we analyze midrank words on the synchronic and diachronic axes, based on time-series and register-varied modern Japanese corpora, respectively. Thus, the distributional properties of mid-rank words can broadly be compared to those of high- and low-rank words under various conditions.

Time-series data comprising n-grams sampled from blog posts is used to examine the role of mid-rank words in detecting rumor trends. We use Shewart's control charts method of identifying abnormal variations in time series data on n-grams with average TF-IDF scores. Having identified mid-rank words having abnormal frequency spikes, we use a word list classified according to semantic principles (*bunruigoihyou*) to uncover collocational patterns in time. For example, the frequency of the mid-rank word "America", which is otherwise a common word, was observed to spike around October 2008, which roughly corresponds to the period when the U.S. subprime mortgage crisis started to unfold. By observing the changes in collocations before, during, and after the frequency spike, it is possible to quantify what categories of words lead up to such a spike.

The Balanced Corpus of Contemporary Written Japanese is used to examine the role of mid-rank word collocation networks in the description of register differences. While common methods in corpus linguistics use keywords, which often correspond to words with a high TF-IDF, or function words, which often correspond to words with a low TF-IDF, to classify the register of documents, we focus on the distributional differences of mid-rank words in predicting register. We show that mid-rank words are less sensitive to specific topics or functional word usage, and can explain aspects of variation not discernible with topic or function words alone.

In conclusion, we show that mid-ranked words are crucial for a comprehensive account of any word or collocation, especially in the frame of thesaurus and collocation dictionary construction. We also identify areas for further research on the viability of mid-rank words in diachronic and synchronic studies, such as the need for more fine-grained classification of the mid-rank.

[Key Words: Corpus Linguistics, TF-IDF, diachronic analysis, Synchronic analysis, register]

JADH 2013 & DH-JAC 2013 CONFERENCE ABSTRACTS

Published by the International Institute for Digital Humanities, Tokyo, Japan

ISBN 978-4-9906708-3-2