A Quantitative Analysis of the Syntax of Classical Japanese Poetry

Kenneth Hanson

2/17/2014 Abstract

Many forms of poetry have an intrinsically quantitative element to them: meter. The constraint of meter forces the author to use various techniques to fit the content of the poem to the form, including the choice of syntactic structures. This study examines these choices in the *tanka*, the traditional Japanese short poem, which consists of five units with a (relatively) consistent 5-7-5-7-7 mora pattern (a mora is a kind of short syllable composing the basic time unit in Japanese). This restrictive form makes an excellent object of study since we can quantify the distribution of different syntactic categories by line and syllable. This study utilizes computer software to parse a text, produce visualizations of the data uncovering patterns that might otherwise be difficult to detect, and use statistics to draw conclusions about these patterns. For the text, I selected *Ogura Hyakunin Isshu*, a Japanese poetry anthology from the Heian Period (c. 800-1200 CE) consisting of one hundred tanka. I parsed the text using the morphological analyzer Mecab and UniDic for Early Middle Japanese, creating a spreadsheet containing part-of-speech tags and Kana (phonetic character) readings for each morpheme in the text. Next, I used a custom-written Python script to add the line number and start/end syllable indices of each morpheme. Finally, I used R statistical programming language to test for correlations between line/syllable position and various syntactic structures. Through this work, I hope to demonstrate the usefulness of methods of corpus linguistics to the understanding of the humanities.