Who writes of Kaguya-hime? Investigating the authorship of *The Tale of the Bamboo Cutter* using stylometric software

J. Cooper Roberts[1], Michael Fang[1], Noah Trudel[1], Patrick Juola[2]
[1]Boston University, [2]Duquesne University

The Tale of the Bamboo Cutter (Japanese: 竹取物語 *taketori monogatari*) is thought to be the oldest surviving piece of classical Japanese fiction. The plot follows an unusual girl named Kaguya, who was found in a hollow bamboo stalk by an elderly bamboo cutter. (Keene 1956). The text has captured the interest of not only scholars but also the general public of Japan for many generations. This can be seen through the wealth of critical scholarship and the plethora of adaptations and references in popular media.

The identity of the original author has, unfortunately, been lost to time. However, there are several hypothesized authors, including the poets Minamoto no Shitagou, Henjou, Ki no Haseo, a member of the Imbe clan, and a member of the faction opposed to Emperor Tenmu (Keene 1999). This is an ideal scenario for the field of authorship attribution, which pertains to identifying the author of an anonymous document from a pool of potential authors. Our objective was to shed light on the matter of who wrote (or at the very least, who did not write) *The Tale of the Bamboo Cutter* using pre-existing methods for authorship attribution.

Authorship attribution for Japanese texts is a particularly challenging task. Without accounting for archaic or uncommon characters, typed English has well under 100 Unicode characters. The same can be said for most languages that utilize an alphabet. Given that Japanese uses a combination syllabary-logograph writing system, its Unicode character count is several thousand. There is the additional problem that typed Japanese does not make use of white space, which is often used by text analysis programs to identify word boundaries (a critical component for some stylometric methods).

Apart from all these issues, it is problematic to assume that stylometric techniques designed for Modern Japanese will yield accurate results for Classical Japanese. It is well understood that written and spoken language change over time. Since the Heian period, Japanese has undergone sound changes, experienced semantic shifts, dropped certain characters, and lost particles, among other things (Shirane 2005). Thus, methods of authorship attribution designed for modern varieties of the language were not utilized. While a model for Classical Japanese is not unprecedented (Tsuchiyama & Murakami 2013), we experimented with several quantitative methods to determine which was most accurate.

We began by constructing a corpus of machine-readable Heian period literature organized by author. Text files were taken from the University of Virginia Library's Japanese Text Initiative.

The corpus included the theorized authors Henjou and Minamoto no Shitagou, as well as seven contemporaneous distractor authors: Fujiwara no Teika, Izumi Shikibu, Kiyohara no Motosuke, the mother of Michitsuna, Murasaki Shikibu, Ono no Komachi, and Sei Shounagon. Ki no Haseo was not included in the corpus because we failed to find any literature from the poet originally written in Japanese. Furthermore, we did not include any member of the Imbe clan or the faction opposed to Tenmu. While theorizing that a member of either party is perfectly valid, the identifier is much too vague for an authorship attribution study.

We decided to include three different genres of Classical Japanese literature, namely *waka* (poetry), *nikki* (diaries), and *monogatari* (tales). Under other circumstances, this mixing of genres in a corpus for authorship attribution should be avoided (Rudman 1997). There are legitimate concerns that this practice is akin to comparing apples to oranges; it is possible that authorship attribution software will distinguish two works by the style of their genre, as opposed to the style of their authors. However, given the fact that Classical Japanese texts (and machine-readable texts in particular) are extremely limited in supply, this was a necessary approach.

Upon completion of the corpus, we combined smaller *waka* into larger files and divided the longer *nikki* and *monogatari* into smaller files. This left us with 252 files in total. We then used the Java Graphical Authorship Attribution Program, or JGAAP (Juola 2009), to find the most accurate method of identifying the author of a document written in Heian period Japanese. Specifically, we experimented with several permutations of the *character n-grams*[1] setting and the *Leave One Out Centroid*[2] analysis method. We chose this method because n-gram methods have been used to great success in authorship attribution studies in a huge variety of languages, including Modern Japanese (Matsuura & Kanada 2001). As illustrated in *fig. 1*, the trigram and 5-gram settings had the highest accuracy[3] (~0.93).

[1] A sequence of *n* successive units, in this case characters.
[2] Compares every text with every other text to determine which author's style the target text is most similar to.
[3] For the purposes of this study, we define accuracy as the number of correct attributions divided by 252, the total number of texts. An example of a correct attribution is as follows: the software posits that Author A wrote Text X over Author B, and Author A wrote Text X.

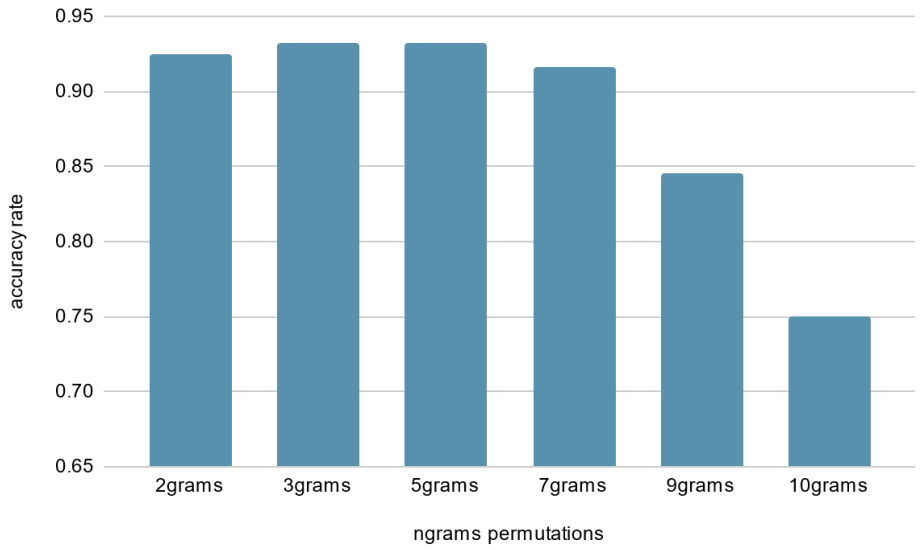

Figure 1: A chart representing the correct attribution rates of various n-gram settings.

We then split *The Tale of the Bamboo Cutter* text into ten smaller pieces. We compared these snippets to the writings of the authors in the corpus using the aforementioned trigram setting in JGAAP. The program attributed seven of the snippets to Sei Shounagon, two of them to Murasaki Shikibu, and one of them to Izumi Shikibu; all three were distractor authors. Consequently, we have found no evidence to suggest that either Henjou or Minamoto was the original author behind the text.

It is curious, though, that all three of these writers were women. Is it possible that the original author of the text was female? We ran a similar experiment to explore this idea, organizing the texts by author gender instead of author identity. Bigram and trigram settings both yielded an accuracy rate of ~0.9841 in training. When we compared the *Tale of the Bamboo Cutter* snippets to our gender-sorted corpus, JGAAP attributed every snippet to the female category. This suggests that the author was female. In the wake of this finding, we believe that a re-examination of other forensic and historical evidence pointing to the tale's author is in order.

Quantitative authorship attribution is not necessarily the silver bullet that will unravel all of the mysteries surrounding this text. However, we are confident that the products of this study will bring us closer to the truth. Our results indicate that neither Minamoto no Shitagou nor Henjou were responsible for *The Tale of the Bamboo Cutter*. Furthermore, it may even be that the author was a woman. It is our hope that these findings will be used in conjunction with other forensic and historical evidence to more accurately posit the author's identity. Furthermore, we hope the model that we have developed for Classical Japanese authorship attribution is utilized for similar cases.

Juola, Patrick. 2006. Authorship Attribution. In *Foundations and Trends in Information Retrieval, Volume 1, Number 3*.

Juola, Patrick. 2009. JGAAP: A System for Comparative Evaluation of Authorship Attribution. In *Division of the Humanities at University of Chicago*.

Keene, Donald. (1999). *Seeds in the heart: Japanese literature from earliest times to the late sixteenth century*. New York: Columbia University Press.

Keene, Donald. (1956). The Tale of the Bamboo Cutter. *Monumenta Nipponica*, *11*(4), 329–355. https://doi.org/10.2307/2382982

Matsuura, Tsukasa, & Yasumasa Kanada. (2000). Extraction of Authors' Characteristics from Japanese Modern Sentences via N-gram Distribution. In *Arikawa S., Morishita S. (eds) Discovery Science. DS 2000. Lecture Notes in Computer Science.* 1967, 315-319. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44418-1_38

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, *31*(4), 351-365.

Shirane, H. (2005). *Classical Japanese: a grammar*. Columbia University Press.

Tsuchiyama, G., & Murakami, M. (2013). Authorship identification of classical Japanese literature using quantitative analysis. *Journal of Mathematics and System Science*, *3*(12), 631.