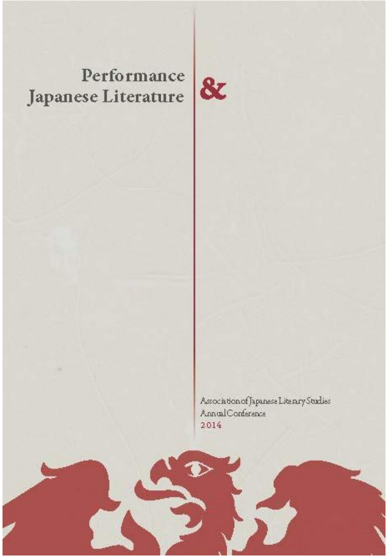
“Reading *The Tale of Genji* through Word Clouds: A Digital Approach to Literary Criticism and Pedagogy”

Catherine Ryu

Proceedings of the Association for Japanese Literary Studies 15 (2014): 58–68.



PAJLS 15:
Performance and Japanese Literature.
Ed. Michael Bourdaghs, Hoyt Long, and Reginald Jackson

Reading *The Tale of Genji* through Word Clouds: A Digital Approach to Literary Criticism and Pedagogy

Catherine Ryu
Michigan State University

This study is an initial attempt to integrate emergent digital technologies into existing critical and pedagogical practices in the field of Japanese literary and cultural studies. In particular, by utilizing Cirrus, one of the computational applications available through Voyant Tools—an open source web-based analytical program (http://voyant-tools.org/)—this study showcases an example computational analysis of one of the three most widely read English translations of *The Tale of Genji*. Ultimately, this study gestures toward the potential of "word clouds" generated by Cirrus as not only a computational tool, but an effective analytical mediator that can loosen the existing disciplinary habits of thought and practice that separate quantitative analysis for the social sciences from qualitative analysis for the humanities.

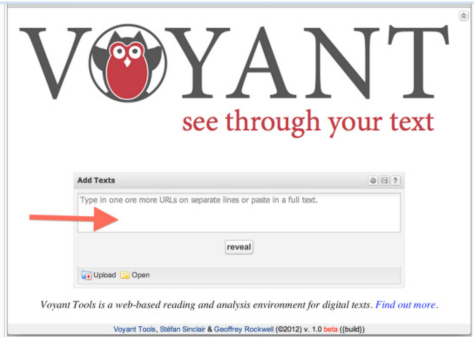
Similar to natural clouds whose shapes, sizes, and colors provide visual cues of climate conditions, word clouds, too, visually represent the verbal climate of a given text. Word clouds render visible a cluster of words used in a text in order of their frequency. That is to say, the higher the frequency, the larger the size of a word in a cloud formation. To generate such a formation, two elements need to come together under optimal conditions: (1) a *free* software program to undertake a computational analysis and (2) *free* digitized texts.

I. How to Generate Word Clouds

Visit the Voyant Tools website and Click on the tools index and choose "Cirrus" (http://voyant-tools.org/tool/Cirrus/). Then, upload a digital text to the "Add Texts" box (Fig. 1). The text can be a plain text or PDF file, or even a URL. A free digital version of Edward Seidensticker's 1987 translation of *The Tale of Genji* is available in its entirety through the Oxford University Computing Services (http://ota.ahds.ac.uk/headers/2245.xml). For the sake of simplicity, this sample demonstration will use a plain text file of Chapter 1 created separated out from this complete version. After inserting the text file into the "Add Texts" box, hit the reveal button. Then a beautiful word cloud emerges (Fig. 2). For the remainder of this study, this word cloud is identified as Word Cloud A to differentiate it from other word clouds to be generated later.

II. How to Collect Word Cloud Data

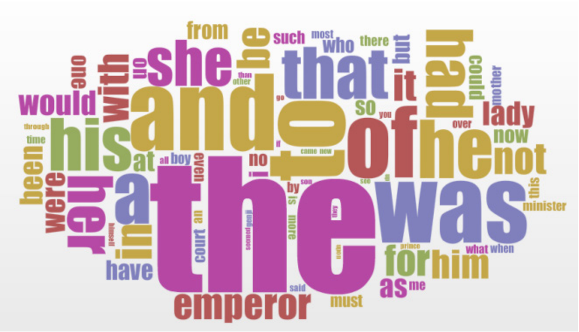
This initial Word Cloud A already tells a lot about the verbal climate of Chapter 1 in *The Tale of Genji* as captured in Seidensticker's English translation. To begin with, some words in the cloud appear larger than others, thereby


VOYANT
see through your text
Add Texts
Type in one ore more URLs on separate lines or paste in a full text.
reveal
Upload
Open

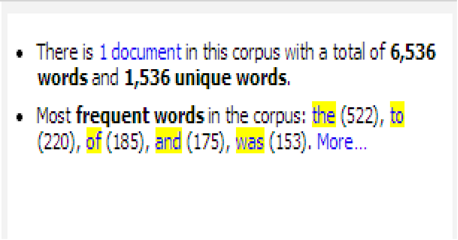
Voyant Tools is a web-based reading and analysis environment for digital texts. Find out more.

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2012) v. 1.0 beta ({build})

(Fig. 1)

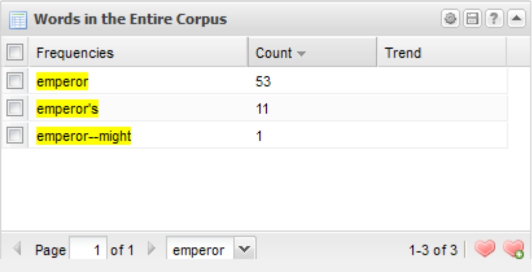


(Fig. 2)



- There is 1 document in this corpus with a total of **6,536 words** and **1,536 unique words**.
- Most **frequent words** in the corpus: the (522), to (220), of (185), and (175), was (153). More...

(Fig. 3)



Words in the Entire Corpus

□	Frequencies	Count ▾	Trend
□	emperor	53	
□	emperor's	11	
□	emperor--might	1	

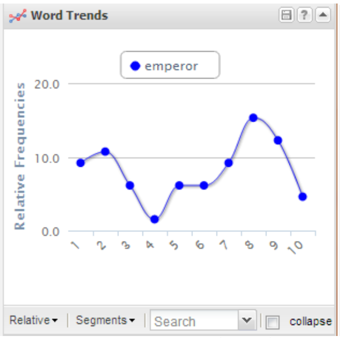
(Fig. 4)

visually representing their relative frequency in this text. A more complete picture of this cloud's makeup emerges when "words" are separated from "unique words," two key terms specific to the realm of word clouds. The former refers to the raw number of words found in a given text or corpus. The latter, unique words, refers specifically to how many unique words the text or corpus contains. As shown in Figure 3 below, Chapter 1 of *The Tale of Genji* is comprised of 6,536 total words, and 1,536 "unique words," meaning that they recur more than once. A list of unique words in order of frequency can also be accessed by clicking the word "more" that appears at the end of the last line in Fig. 3.

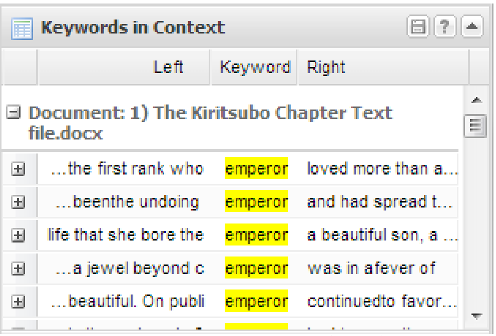
The data tied to the frequently used word list above represents more than numbers. Given that the text used for this study is written in English, it is not surprising that the definite article "the" appears with the highest frequency (522 times). But it is meaningful to note, for example, that the very first verb that appears on this list is the verb of existence; namely, to "be" but in the past tense: "was" is used 153 times. This underscores the fact that the story is not only told in the retrospective but also that the story itself focuses largely on states of being rather than on active motion or movement.

The frequently used word data can reveal an important dimension of the text that unveils itself after the layer of functional terms in English, such as "the," without any unique semantic meaning, is removed. Significantly, if one scrolls down the list of unique words, the first noun that eventually appears on the list is the term "emperor." In fact, this term appears in Word Cloud A (Fig. 2) relatively larger than other nouns such as "court" and "lady." The data generated by this word cloud thus identifies the emperor as the main character in this opening chapter of *The Tale of Genji*. Moreover, since this sovereign is not identified by his personal name, we can surmise that it is his imperial role that is emphasized in this English translation. In fact, a fuller picture of the weight given to the term "emperor" emerges through another search that focuses on the frequency of associated terms. In the case of the term "emperor," its related word is "emperor's," and together there are sixty-four occurrences as shown in Figure 4.

Another important advantage of Cirrus as tool for computational analysis is that it can generate graphs of word frequency over the length of a text, making it possible to visually convey the locations of the term "emperor" in Chapter 1. The graph below (Fig. 5) shows the frequency of the term in the text, which is divided into ten equal segments. The frequency peaks toward the end of the chapter, around Segment #8. Still another set of data can be accessed in the section called "key words in context [KWIC]" (Fig. 6). The same information can be obtained by clicking on the word emperor in the section called "corpus reader," in the middle of the screen, which allows the user to view the full text (Fig. 7).

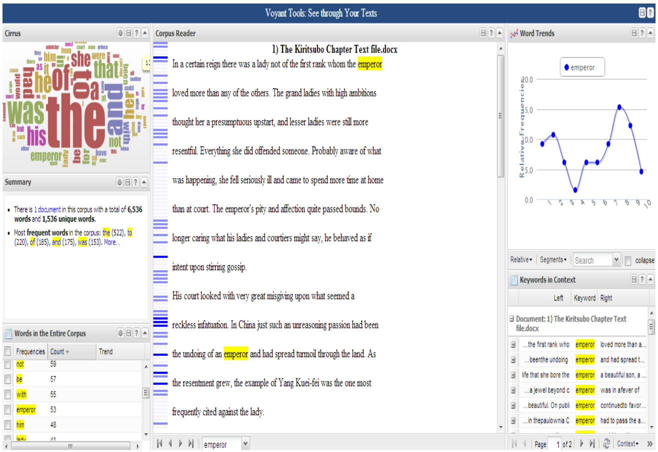


(Fig. 5)


Keywords in Context

	Left	Keyword	Right
⊟	Document: 1) The Kiritsubo Chapter Text file.docx		
⊞	...the first rank who	emperor	loved more than a...
⊞	...beenthe undoing	emperor	and had spread t...
⊞	life that she bore the	emperor	a beautiful son, a ...
⊞	...a jewel beyond c	emperor	was in afever of
⊞	...beautiful. On publi	emperor	continuedto favor...

(Fig. 6)



Summary

- There is 1 document in this corpus with a total of **6,536 words** and **1,536 unique words.**
- Most **frequent words** in the corpus: the (522), to (220), of (185), and (175), was (153). More...

Words in the Entire Corpus

	Frequencies	Count	Trend
□	not	59	
□	be	57	
□	with	55	
□	emperor	53	
□	him	48	
□	lady	41	

Corpus Reader

1) The Kiritsubo Chapter Text file.docx

In a certain reign there was a lady not of the first rank whom the emperor

loved more than any of the others. The grand ladies with high ambitions

thought her a presumptuous upstart, and lesser ladies were still more

resentful. Everything she did offended someone. Probably aware of what

was happening, she fell seriously ill and came to spend more time at home

than at court. The emperor's pity and affection quite passed bounds. No

longer caring what his ladies and courtiers might say, he behaved as if

intent upon stirring gossip.

His court looked with very great misgiving upon what seemed a

reckless infatuation. In China just such an unreasoning passion had been

the undoing of an emperor and had spread turmoil through the land. As

the resentment grew, the example of Yang Kuei-fei was the one most

frequently cited against the lady.

emperor

Keywords in Context

	Left	Keyword	Right
⊟	Document: 1) The Kiritsubo Chapter Text file.docx		
⊞	...the first rank who	emperor	loved more than a...
⊞	...beenthe undoing	emperor	and had spread t...
⊞	life that she bore the	emperor	a beautiful son, a ...
⊞	...a jewel beyond c	emperor	was in afever of
⊞	...beautiful. On publi	emperor	continuedto favor...
⊞	...in thepaulownia C	emperor	had to pass the a...

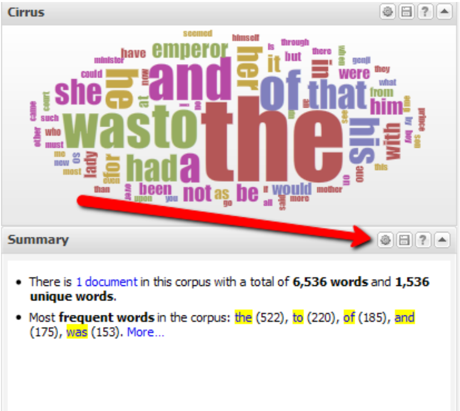
Page 1 of 2 Context

(Fig. 7)

III. How to Fine-Tune Word Cloud Data

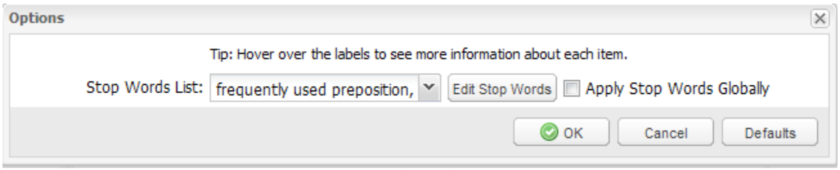
The beauty of word clouds as an analytical tool is that they can be tailored to satisfy one's own pedagogical or research needs. This is perhaps the most exciting and potent aspect of word clouds and dispels a commonly held misconception that computational analysis can only be a blunt instrument of analysis when applied to literary texts. Key to effectively utilizing this potential is the function called "stop words." Stop words refer to any terms not to be included in the formation of a word cloud. For instance, the definite particle "the" can be taken out from the word cloud. To do that, click on the wheel icon located on the lower right corner of the word cloud screen.

Then an options window opens up. Select "Edit stop words" and list the words to be excluded. By taking out a set of words that I will identify as "functional terms," it is possible to reveal deeper levels of the text. Functional terms are words that do not carry discrete semantic meanings in the English language but are functionally necessary. Some examples of functional terms include prepositions, pronouns, and articles. In fact, functional terms usually appear with the greatest frequency in any given text. After creating a list of functional terms as stop words, remember to save that list with a clearly identifiable name so that it can be modified to create word clouds of yet deeper levels of the text. After saving the list, click on the button, “apply stop words globally” to generate a new word cloud formation: Word Cloud B (Fig. 10).



- There is 1 document in this corpus with a total of **6,536 words** and **1,536 unique words**.
- Most **frequent words** in the corpus: the (522), to (220), of (185), and (175), was (153). More...

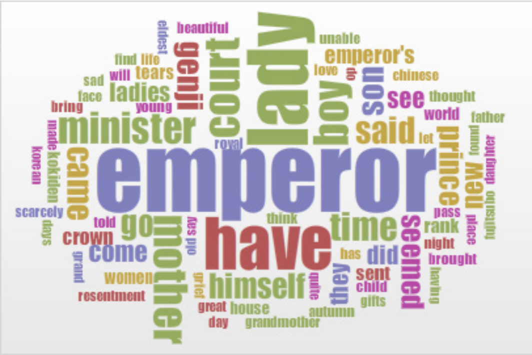
(Fig. 8)



Tip: Hover over the labels to see more information about each item.

Stop Words List: frequently used preposition, □ Apply Stop Words Globally

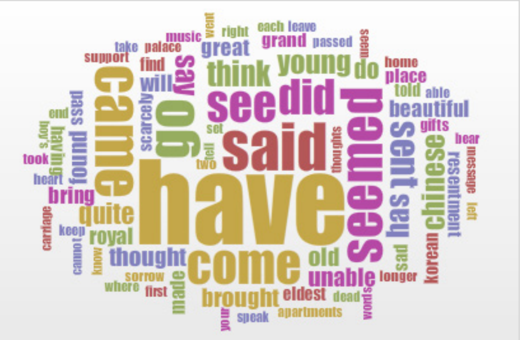
(Fig. 9)



(Fig. 10)

Word Cloud B reveals a deeper layer of the text, now with semantic values specific to the first chapter of *The Tale of Genji*. Not surprisingly, "emperor" is the most frequently used term, along with two other common nouns—"lady" and "minister." This visualized data brings into focus that the emperor's main relationships pertain both to his private and public lives as represented by two key terms. What is emphasized here is, again, these characters' roles rather than their individual identities. A closer look at this second word cloud reveals such terms as "palace," "court," "prince," "royal," "resentment," "support," "music," etc.—these key words now paint a more refined picture of what is recognizably the world of Genji.

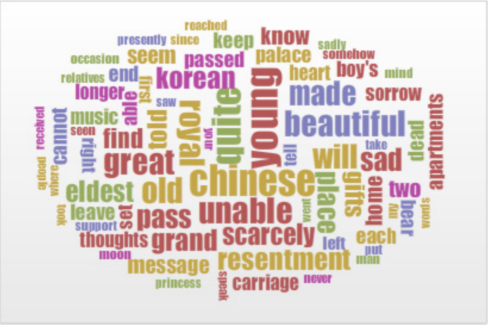
By taking out the most frequently used nouns related to key characters such as "emperor," "lady," "minister," etc., Word Cloud C (Fig. 11) reveals a yet deeper level of the world of Genji. This is the world that is made largely of verbs and adjectives, with verbs appearing with greater frequency than adjectives.



(Fig. 11)

Significantly, other than the verb "to be," the actions most commonly described in the first chapter of *The Tale of Genji* are those of possessing, seeing, coming and going, speaking, and thinking. These are actions pertaining to observation, communication, and rumination. This information can potentially illuminate both the specific process of how the characters, in the first chapter of *The Tale of Genji*, develop their relationships with one another, and the nature of their relationship.

By further removing the verbs appearing with higher frequencies, a new word cloud (Fig. 12) is formed, revealing yet another aspect of the world of Genji.

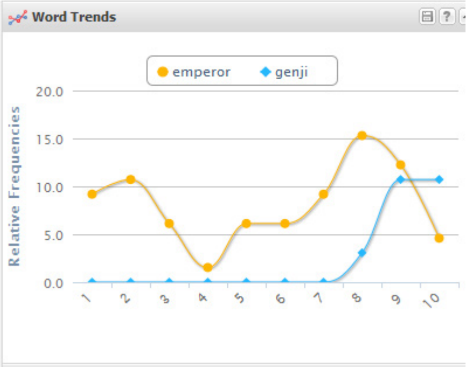


(Fig. 12)

In this Word Cloud D, key adjectives have become more pronounced than before. The combined forces of such adjectives as "sad," "great," "grand," "the eldest," "royal," "beautiful," etc., paint the emotional tenor of this first chapter. In light of the earlier set of most frequently used verbs that do not focus on vigorous actions, together with this set of adjectives and emotive words, it now becomes relatively easy to explain why *The Tale of Genji* has been appreciated as a psychological novel. By applying a series of stop words lists, it is possible to obtain finely tuned statistical data about a given text, both at the micro and macro levels, and with a specific interpretive focus.

IV. How to Utilize Word Clouds as a Tool for Literary Analysis and Pedagogy

Word clouds can be a powerful pedagogical tool, especially if used to instruct uninitiated readers of literature. There is a widely shared myth that only a certain type of people can read, understand, and analyze literature. Word clouds can demystify this myth. For example, the graph below (Fig. 13) shows the frequency of the terms "emperor" and "Genji" together.

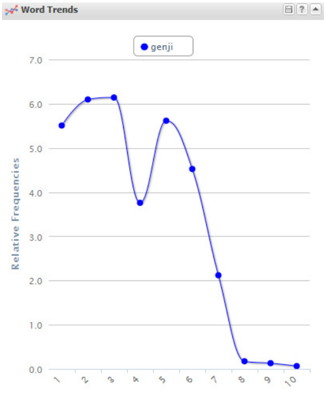


(Fig. 13)

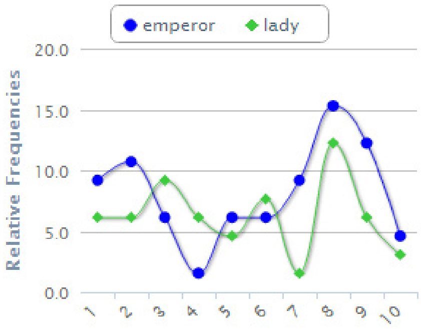
This graph can be used to persuade students that by the end of the first chapter, Genji has emerged as the main focus of this narrative since his proper name appears more frequently than the term "emperor." After showing this graph and explaining how to interpret it, we can ask a set of related questions, or ask students to generate their own questions, regarding the relationship between Genji and the emperor, his father.

For example, "Why is it important to focus mainly on the emperor in Chapter 1, when Genji is the eponymous hero of this tale?" This graph can also be paired with another one that shows the frequency of the term "Genji" over the entire text (Fig. 14). Here we see the frequency of Genji's name dropping dramatically around Segment 7, after he dies in Chapter 41, although it does not disappear completely. Why does the story go on after the hero's death, and how, and why, does his name reappear in the remainder of the story?

Another approach is to use a graph showing that in Chapter 1, the term "emperor" frequently appears in close proximity to another term, "lady" (Fig. 15). When we examine the lady in question by investigating each occurrence of the term in context, this term refers to the emperor's three main ladies—one is Genji's mother, who passed away when he was an infant, and the other two are his stepmothers. Given that one of the stepmothers is his archenemy and the other stepmother is the object of Genji's own desire, what kind of relationship can Genji possibly develop with his father? Asking such questions based on the data generated by word clouds may be an effective way of getting students intrigued about *The Tale of Genji* and thinking more deeply on their own about this story.



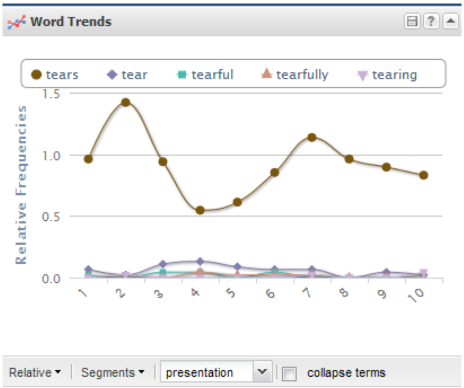
(Fig. 14)



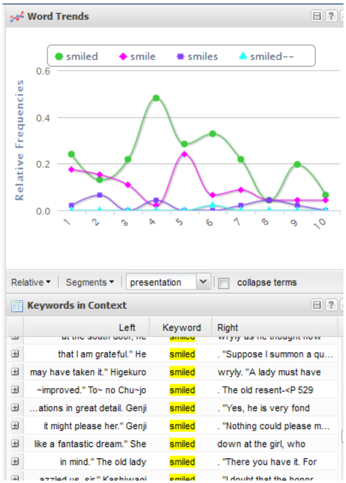
(Fig. 15)

Even for scholars and critics, word clouds can serve as a powerful tool, especially when undertaking a large-scale research project requiring both quantitative and qualitative analyses. For example, *The Tale of Genji* has been celebrated as the embodiment of Japanese aesthetic sensibility, *aware* (the pathos of things), but it is possible to evaluate such a claim empirically with hard data by generating the patterns and frequencies of emotive words used not only in Chapter 1 but in the entirety of *The Tale of Genji*. As expected, a set of such associated words as "sadness," "tears," and "regret" appear throughout the narrative, creating the overall nuance of the story one of *aware* (Fig. 16).

Yet, surprisingly, in the entirety of *The Tale of Genji*, a set of words related to "smile" also appears with a higher frequency than anticipated (Fig. 17). A perusal through the word in context reveals that Genji turns out to be the champion of smiles, especially in a later stage of his life when he is basking in glory and plays with the fate of other characters, such as Tamakazura. Besides Genji, however, other people are also described smiling on various occasions, thereby altering the received notion of *The Tale of Genji* largely and mainly as an embodiment of *aware*.



(Fig. 16)

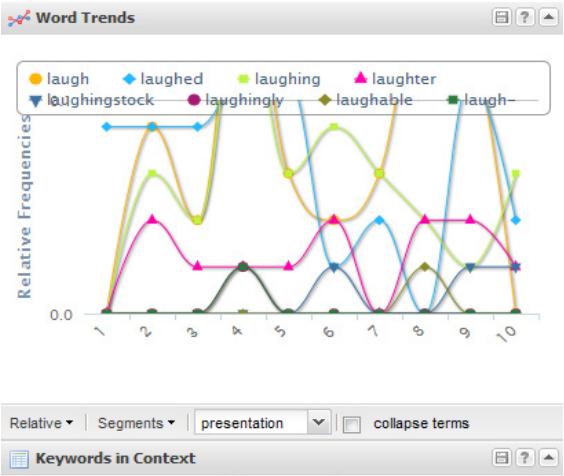


Left	Keyword	Right
at the south door, he	smiled	wryly as he thought how
that I am grateful." He	smiled	. "Suppose I summon a qu...
may have taken it." Higekuro	smiled	wryly. "A lady must have
~improved." To~ no Chu~jo	smiled	. The old resent-<P 529
...ations in great detail. Genji	smiled	. "Yes, he is very fond
it might please her." Genji	smiled	. "Nothing could please m...
like a fantastic dream." She	smiled	down at the girl, who
in mind." The old lady	smiled	. "There you have it. For
azzled us, sir." Kashiwagi	smiled	. "I doubt that the honor

(Fig. 17)

What is perhaps even more surprising is the frequency of another set of words related to "laughing" (Fig. 18). But when examined in context, this open expression of joviality occurs paired with "weeping and laughing," or to describe a baby's reactions, or to express how certain characters are rather afraid of "being laughed at." *The Tale of Genji* does not embody the Japanese aesthetic sensibility in an abstract sense but in a concrete social context. In other words, with finely calibrated computational data and interpretative strategies, it is possible generate a compelling reading of *The Tale of Genji* both at the micro and macro levels simultaneously.

The ultimate significance of such a reading can be further analyzed in light of the original text in classical Japanese. The differences and similarities that emerge from this project would be an important case study of the cultural politics in translation studies, for instance. As this brief study has shown,



(Fig. 18)

quantitative data generated through word clouds can transform the misguided perception that literary analysis is merely a subjective interpretation without substantial evidence.

What is most exciting about word clouds, however, is that they are not merely a tool of analysis but an embodiment of frequency-based analysis in action. As a matter fact, word lists, such as the ones created by the word clouds used in this study, are increasingly used to generate new forms of knowledge and practice in such fields as corpus linguistics, frequency-based English grammar, digital humanities, etc.[1] With available and emergent technologies that utilize frequency-based analysis, it is possible to create an alternate approach to teaching classical Japanese grammar in particular. Instead of introducing students first to the daunting conjugation tables of nine different types of verbs, each with six inflected forms, as has been a common practice, we can lead them to the exciting verbal realm of classical Japanese through a word list. Such a list can be comprised of two or three units of most frequently used expressions, given in their lexical context as they appear in different genres of classical Japanese texts. In this way, students can learn to recognize the most frequently used patterns of expression, rather than the individual components that make up classical Japanese grammar.

[1] For a glimpse into how frequency-based approaches have been applied to second language teaching and research in the field of linguistics, refer to Douglas Biber and Randi Reppen, "What Does Frequency Have to Do with Grammar Teaching?," *Studies in Second Language Acquisition* 24.2 (2002): 199-208; Minju Kim, "Discourse, Frequency, and the Emergence of Grammar: A Corpus-based Study of the Grammaticalization of the Korean Existential Verb is (i)-ta" (PhD Diss., University of California Los Angeles, 2003); Joan Bybee, "Usage-Based Grammar and Second Language Acquisition," in *Handbook of Cognitive Linguistics and Second Language Acquisition*, eds. Peter Robinson and Nick C. Ellis (New York: Routledge, 2008), 216-36.

Moreover, once a set of the most frequently used parts of speech or verbal patterns is identified, they can be used in games of pattern recognition, further opening up a new space of learning and teaching.[2] In other words, a creative use of currently available digital technologies can potentially and dramatically transform existing critical and pedagogical practices in the field of Japanese literary and cultural studies. It can be carried out by integrating computational analysis with close textual reading practice, by using corpus linguistics to create a "gamified" approach to learning based upon skills of pattern recognition, or by ways that have hitherto been unexplored and unimagined. This study thus gestures toward the future of Japanese literary and cultural studies in an increasingly digitized environment.

[2] I am, in fact, in the process of creating such a language-learning game platform to teach classical Japanese grammar. This platform is called Cube2Cube (C2C) and is U.S. patent pending, filed by Michigan State University Technologies (March 11, 2013).