Harshita Gupta

Humanities Colloqium

April 26, 2017

Cover Letter

In developing this revision, I used Franco Moretti's "Network Theory, Plot Analysis" from *Distant Reading* as a model of a work in a similar genre. The day after our conference, I was finally able to get my network code to work, and figured out how to visualize topics as a graph connecting different words and elements across texts. This gave me a very productive experimental setup for understanding what I'd been interested in all along, the unique thematic 'structure' of each of the texts. As Moretti says in his analysis of Hamlet through a network of its plot,

> Finally - and it is the most important thing of all, but also the most difficult - one can intervene on a model; make experiments. Take the protagonist again. For literary critics, this figure is important because it is a very meaningful part of the text; there is always a lot to be said about it; we would never think of discussing Hamletwithout Hamlet. But this is exactly what network theory tempts us to do: take the Hamlet-network, and remove Hamlet, to see what happens.

I engage in this process of "making experiments" on the networks that I produce, by deleting different variations of words like 'women', 'daughter', 'mother', and 'mistress', and subsequently understanding their effect on the network: to put it differently: I attempt to understand each translation with questions like "What remains of the of each translation's depiction of women when we remove mentions of all male - associated women, like mothers, wives, and daughters?" "Do women remain in the networks, and to what extent, when we remove mentions of courtly duties and performance, versus when we remove mentions of feeling, emotion, and thought?" How does each translation's network respond differently to these interventions? Engaging with the models in this way was the most challenging part of HUM10 all year, since it was unlike anything I've done before; it was very difficult to think of productive questions that would reveal interesting trends, write code to answer the questions, wait for it to run, and analyze my results thoughtfully without getting lost in the 80 or so files that each run of code would generate.

When I cleaned the models according to the interests I expressed in draft, removing some rogue proper nouns and publishers notes, many of the trends that I'd identified in my draft, like the difference in occurrences of women, were no longer present. Generating the models with the parts of speech that I detected earlier, and with the topic number setting set to 15, also yielded cleaner, more meaningful topics that didn't have as many unreadable words or errors. Additionally, instead of simply using the top ten words in a topic, I used a threshold value to determine whether the topics coherence was statistically significant (threshold signal value); I felt

like this was a representation that relied on a more rigorous understanding of the model's output.

I think much of the difficulty with this paper was that I'm not as well trained in reading and interpreting models as I am in close-reading — setting out on this project was a challenge in both, the model-creation respect, as well as in learning to productively and accurately interpret topic signals and force-directed network graphs.

Given more time and resources, I would certainly give these models and their results a more thorough treatment, with perhaps a more "mesoanalytic" lens — since I wasn't able to read all four translations in depth, and was only interacting with models, the traditionalist in me (however small a part that may be) finds the result only partly satisfying. I would want to endeavor to correlate my network's results with clearer plot and thematic trends across the text, versus ones in specific scenes that the topics can point out; this would only be doable once I've read the entirety of each text.

Computational X-Rays: The Aging Thematic Bones to *Genji's* Translations

In Metonymy in The Tale of Genji: An Analysis of Translation Strategies, Janel R. Goodman Murakami compares occurrences of metonymy in a passage across translations of the Tale of Genji to assess how domesticated or foreign Suematsu's, Waley's, Seidensticker's, and Tyler's translations are, concluding that the more modern translators retain foreign elements of the text more faithfully than their earlier counterparts. In "Going to Bed with Waley: How Murasaki Shikibu does and does not become world literature," Valerie Henitiuk uses a similar microanalytic approach, more commonly referred to as close-reading, to critique Waley's translation of Genji and his portrayal of women. I build on the preexisting body of scholarship on Japanese to English translations of Genji, instead using the macroanalytic approach of topic modeling, to compute the themes across the translations, and analyze what discrepancy between translations reflects about the text's portrayal of women. I ultimately build upon Murakami's conclusion, determining that more recent translations not only portray Heian Japan more faithfully, but also elevate the position of women as individuals without "censor[ing]," as Henitiuk put it, male-female interactions to the same degree as older ones.

I apply digital topic modeling to break down and analyze three translations of Genji: Arthur Waley's from 1925, Edward G. Seidensticker's from 1976, Royall Tyler's from 2001, and Dennis Washburn's from 2015 [1]. Topic modeling is a natural language processing technique that uses the principle of distributional semantics, or the common cooccurrence of two words, to group words into "topics." I use the Latent Dirichlet Allocation (LDA) topic modeling method, which operates on 1000-word sequential chunks of the tale of Genji. LDA treats each "chunk" of

[1]My greatest interest was originally in comparing the work of male translators with Helen McCullough's partial translation. Unfortunately, McCullough's translation, published only in *Genji and Heike: selections from the Tale of Genji and the Tale of the Heike*, is heavily copyrighted and unavailable in any digital format.

words as an entity composed of "topics" —each topic is composed of words that commonly occur together. A statistical explanation of LDA modeling's assumptions and mechanisms is beyond the scope of this paper, and one can turn to Rhody's "Unpacking the Assumptions of LDA" or Jockers' Macroanalysis for a simplified analysis of the method. Crucial to this paper and its discussion, however, is LDA modeling's unsupervised nature. Topics are not produced based on the program's understanding of the words' meanings or potential similarity, but creates buckets for topics based on the position of words relative to each other, and their common cooccurrence, i.e. distributional similarity, to determine that they belong to a similar topic. Its unsupervised nature makes LDA modeling useful for identifying topics in a more "objective" fashion by identifying authors' subconscious placement of words and themes and therefore reflecting their gendered and cultural inclinations. Topic modeling is useful, therefore, not just due to ability to give us a zoomed-out, abstracted view of a text as large as Genji, but also in uncovering topics that might be outside our microanalysis-based understanding of discoverable topics.

Topic modeling is traditionally applied to multiple documents that are dissimilar, as in the work by Jockers, and has been trained on corpuses containing multiple texts, 4500 poems in Rhody's case and 4500 texts in Jockers' case. These models develop generalized, often easily identifiable themes that can apply to texts across time and author, like Rhody's "night light moon stars day dark" and "tree green summer flowers grass" topics. Jockers' and Rhody's topics reveal nothing surprising in their content themselves; the combinations of words are predictable. Topic modeling of the form they practice is useful when the purpose of the topics, albeit predictable, is to be used later in topic distribution comparison across texts, but is less useful when working with a single text. I train the LDA model, instead, on a single translation of Genji at a time, thereby developing topics that are far more specific to each translation's "thematic world," and useful in

Figure 1: Topics in Waley's Translation

Topic 1: said now look even man old come make seem mother tree young hous
Topic 2: said time now letter ladi long even much feel littl last go visit way far
Topic 3: cat udoneri princess palac seen still fine creatur
Topic 4: said go now thing come know think way see seem look say time thought get make much long feel tell peopl back last well day night inde even girl moment hous old felt quit take
Topic 5: ladi side
Topic 6: mistleto rebuilt
Topic 7: year inde cold dress imagin
Topic 8: now time day ladi said much emperor inde father great even way seem princ year thought palac daughter feel made make long matter girl old mother see littl
Topic 9: now ladi time said day even seem thought great long littl look come poem inde
Topic 10: time day now ladi inde much even said seem come felt long thing way feel
Topic 11: now boy place emperor princ hous
Topic 12: emperor palac great present
Topic 13: emperor ladi princess art

identifying the thematic differences underlying each translation. The interpretive advantage to this method is evident in the difference between the word clouds generated by the two cited authors, and the ones I display and analyze in my paper.

Developing meaningful topics is not just a matter of corpus scale, but also of corpus content: following Jockers' guidance, I remove all stop words that do not provide interpretive value, like articles and pronouns, from the corpus before I discover themes within it. In "Theme", Jockers proposes and demonstrates topic modeling only on a corpus of common nouns. Restricting the corpus to nouns is appropriate in his use case as his corpus is much larger than a single *Genji* translation, and therefore filtering for nouns allows for a higher level of abstraction that focuses on broader trends that are more likely to occur across hundreds of texts. In analyzing *Genji*, however, restricting my models to only nouns results in a loss of rich nuance which is central to a cross-translation comparison: adjectives, adverbs, and verbs contain a wealth of information about emotion, description, and subjectivity. To this end, I include adverbs,

Figure 2: Topics in Seidensticker's Translation

Topic 1: stag
Topic 2: seem come said t even see littl bishop dream came child look old girl
Topic 3: seem robe paint red princess ladi women line littl white string master nose now
Topic 4: girl ladi said seem thought think come now t look littl even know daughter good young time see women mother make came much go well man way princ
Topic 5: seem thought even think come see now said go ladi princess time thing littl women make know world long want much day way say made feel look came still noth father
Topic 6: emperor princ thought seem ladi court chines crown daughter present palac time son said year royal
Topic 7: princ ladi emperor said blossom kobai daughter thought carriag even
Topic 8: seem thought women littl governor even young room ladi look light said came back open away door wind way see day veranda
Topic 9: emperor time now matter tear princ minist empress
Topic 10: seem ladi safflow quarter white look
Topic 11: ladi seem thought now time day even said year come see old princess came made much littl
Topic 12: seem ladi thought daughter year come old think said son day thing well great made man
Topic 13: t man thing good governor say seem girl think make look know go happen said woman daughter wife one want young day come even just don ve sort someon see everyth peopl m let ask
Topic 14: emperor ladi court cat princ son crown mother third new father go see royal

Figure 3: Topics in Tyler's Translation

Topic 1: high paper great shoot book write cup scroll command son
Topic 2: now high even look said way well still thought never time see littl go come long know feel seem day just thing ladi
Topic 3: play captain old tale now flute music never ladi feel said seem life day hear say even
Topic 4: majesti now flower light even wind long morn said dew day
Topic 5: blossom look command made flower ladi littl spring beauti year
Topic 6: now monk day world holi long mountain come adept wind littl captain thought citi know rever sent look way time
Topic 7: rever nun young woman lordship go die come mother said even someon tell now noth high see look seem ask never sure just back told heard well happen talk sister women mistress
Topic 8: majesti flower spring autumn ladi nun music year pine east quarter made consort wing long blossom high garden color
Topic 9: captain young excel high play look lieuten son music blossom said right just seem man dress one even well make daughter come ladi secretari advis
Topic 10: even now old play novic high noth often see biwa
Topic 11: ladi gown look dress comb box one never mistress even high far day now cathay plum
Topic 12: now long majesti littl thought day mani ladi made year see said well citi time still even come go look

Figure 4: Topics in Washburn's Translation

Topic 1: go ladi just woman feel come girl know even way said make think now nurs tell ask littl man lord night well look time dont attend see thought place
Topic 2: carriag go feel wife process men view attend ladi just way blind space
Topic 3: feel husband now even ladi emot heart deep woman say
Topic 4: made lotu attend third minist flower make
Topic 5: even capit boat day governor sea attend provinc ladi wind perform lord dream wave
Topic 6: robe look ladi poem even day garden now seem blossom color tree made women
Topic 7: cat third princess princ son crown littl blind robe felt see look close game face never contest think tri remark
Topic 8: even feel now time world look thought see come ladi live heart felt just princess long villa day go
Topic 9: father hardli wife
Topic 10: time daughter son even emperor palac princ princess now father court look day year majesti third minist mani feel ladi world left
Topic 11: play koto string instrument perform princess even music young son boy littl seem flute hear time daughter look major just robe women feel
Topic 12: woman make way even man young just time now thing matter go mani
Topic 13: princess even look feel young daughter woman man just ladi now go think women make time way captain letter thought see thing say come still
Topic 14: ladi wind princess day ceremoni daughter father come carriag

adjectives, and verbs in my topic models.[2] The topics discovered, with the words that have the highest percentage of appearance within a given topic, are included in figures 1, 2, 3, and 4 [3].

The topics discovered in the four translations, reproduced in the figures above, reveal some clear trends across the four translations. If we look simply for the presence of women across texts: 'daughter', 'lady', 'princess', 'wife', 'women', 'woman', 'mother', 'girl', 'sister', and 'mistress', we find that 9 of the 13 themes in Waley's translation include women, 13 of the 14 in

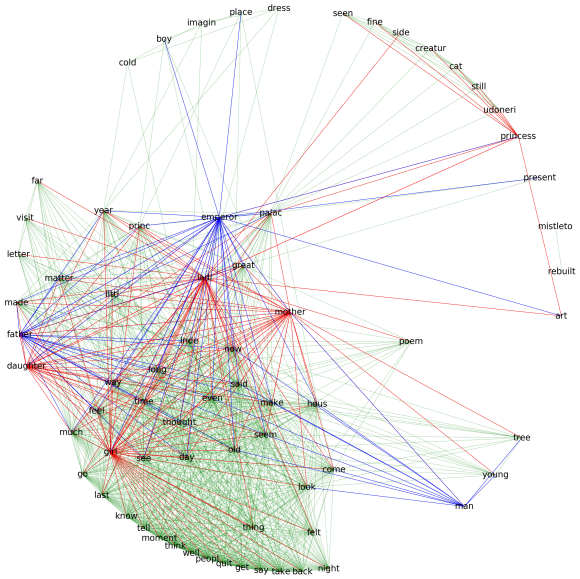
[2]In addition to removing stop words and parts of speech other than the ones specified, I removed all annotations, introductions, publishers' notes, footnotes,and endnotes. In the Tyler translation, this included the deletion of all chapter titles that are not from the original, chapter introductions, and the "Persons" and "Relationship to Previous Chapters" sections. Additionally, all words are reduced to their stems - therefore treating "respect" and "respectfully" as identical semantic units and indistinguishable to the LDA algorithm. I also exclude all proper nouns from the corpus used for the topic model, so that topics are identified not upon the basis of scene (which correlates highly with proper nouns like names and places) but upon common motifs. Special care was taken to exclude all Japanese proper nouns, which are not identified by Western natural language processing tools. For a list of all Japanese proper nouns removed, see the code in Appendix A.

[3]All words in a given topic with signal strength greater than 0.0040 are included, i.e. words that comprise more than four percent of a given topic. If a topic had no words with signal strengths high enough, i.e. the topic was statistically insignificant, those topics were excluded from the final model and not used in subsequent discussion.

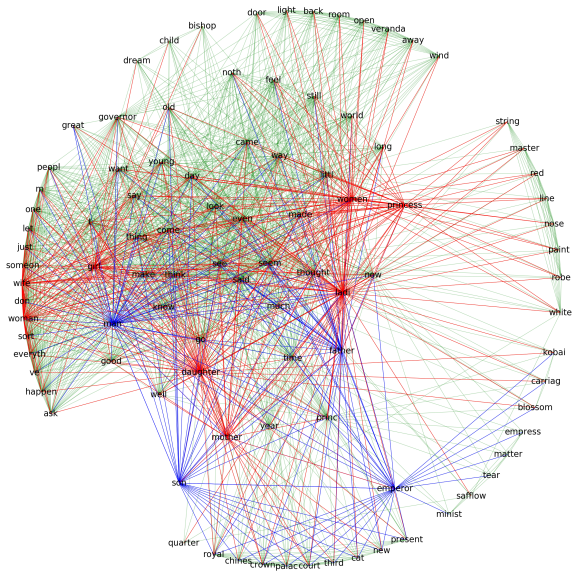
Seidensticker's do, 8 of 12 in Tyler's do, and 13 of 14 in Washburn's do. Mentions of men: 'son', 'prince', 'father', 'emperor', 'minister', 'lord', 'man', and 'husband', on the other hand, are only 5 of 13 in Waley, 8 of 14 in Seidensticker, 2 of 12 in Tyler, and 10 of 14 in Washburn.

Proceeding to analyze this thematic data numerically may be fruitful, but it doesn't provide as much a visual 'picture' of the interactions between topics across texts. In selecting the appropriate visualization mechanism for this experiment, I turn to my original intent: to understand the underlying thematic structure to each translation, and analyze this structure to come to macroanalytic conclusions about the translations in comparison to each other. Networks representing the words each each topic, with edges connecting words that co-occur in a topic, accomplish just that. I take inspiration from Franco Moretti's essay "Network Theory, Plot Analysis", in which he set the precedent for combining network theory and literary analysis; Moretti draws network diagrams to depict interactions between characters in *Hamlet*. While his graphs are hand-drawn, I generate mine programmatically and use force directed graph drawing to display them, a technique which minimizes edge overlap and places appropriately more "central" nodes, i.e. words which have more edges connected to them, and "peripheral" ones, i.e. words which have fewer edges connected to them. Force-directed drawing of the topics in *Genji* reveals immediately the words central to each translation, the clusters of words, and the clustering of topics of words relative to each other. Network diagrams of the four translations considered in this paper are in figure 5, with edges connected to female-associated words drawn in red, edges connected to male-associated words drawn in blue, and all other edges drawn in green.

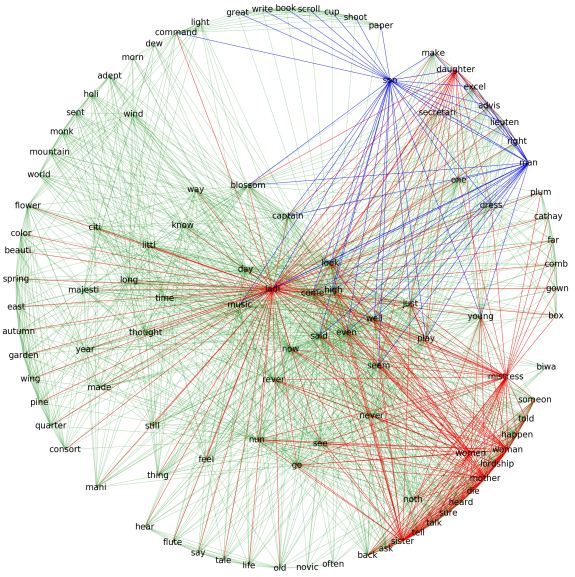
The numerical observations made earlier jump out immediately now: the prevalence of women-associated words across the networks and the higher ratio of red to blue. As Moretti put it,



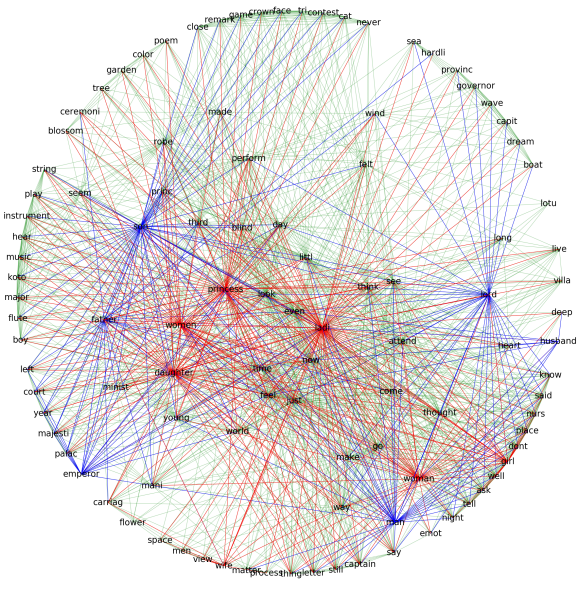
(a) Waley



(b) Seidensticker



(c) Tyler



(d) Washburn

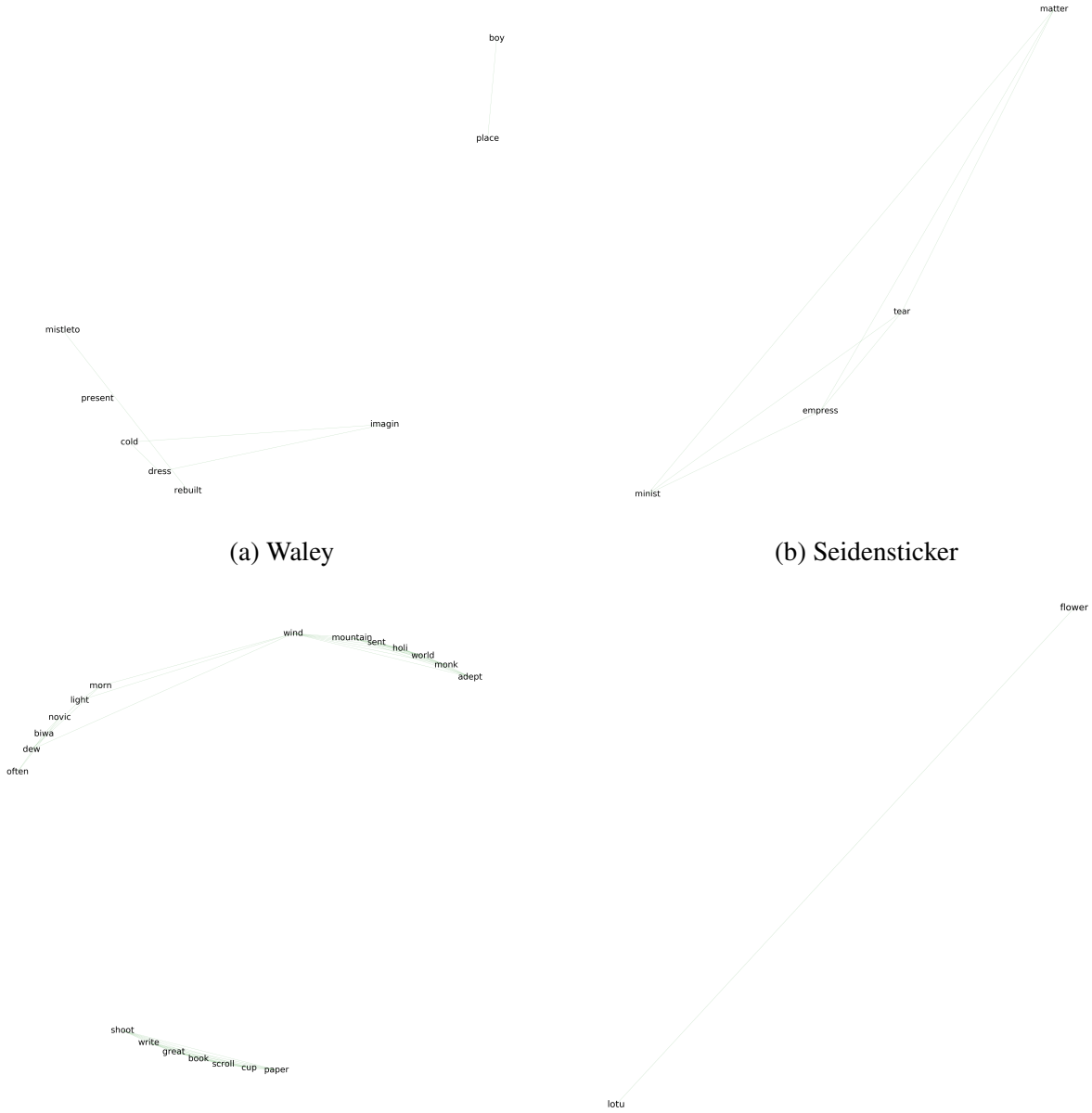
Figure 5: *Genji's* Thematic Structure: Female-Associated Connections in Red, Male-Associated Connections in Blue

> "Once you make a network of a play, you stop working on the play proper, and work on a *model* instead... this process of reduction and abstraction makes the model obviously much less than the original object... but also, in another sense, much *more* than it, because a model allows you to see the underlying structures of a complex object. It's like an X-ray" (Moretti 218).

Beyond what is apparent via color: looking at the centrality of female and male nodes in the graph reveals that Seidensticker's and Washburn's representations of women are far more central to the thematic networks of their translations, with the female nodes having very high degrees (number of edges originating from them) and extending more uniformly across the graph. Tyler's and Waley's, on the other hand, while prevalent, are far more peripheral, suggesting a portrayal of women that is persistent yet a near afterthought, or secondary to the less frequent yet more central male occurrences.

To probe into the dependencies and relationships that characterize this network, I follow Moretti's lead in attempting to "*intervene* on a model; make experiments (Moretti 220)." The first is a simple one for purposes of demonstration: what happens when all women-associated nodes and the nodes that they connect to, are removed from each network, versus when all male-associated nodes are? The results are in figures 6 and 7. The effects is apparent: in each case, the network collapses, unsurprisingly, but perhaps what is more surprising is that the removal of one word group obliterates the other as well – both blue and red in either experiment – showing how pervasive the distributional linkage of men and women in *Genji* is, across translational guidelines.

Examining the non-gendered (as we may currently believe) nodes central to the graphs reveals fruitful direction for future experiments. 'feel' is the most central node in the Washburn graph, with 'thought' being a close second. 'Perform', 'attend', and 'robe' introduce the element



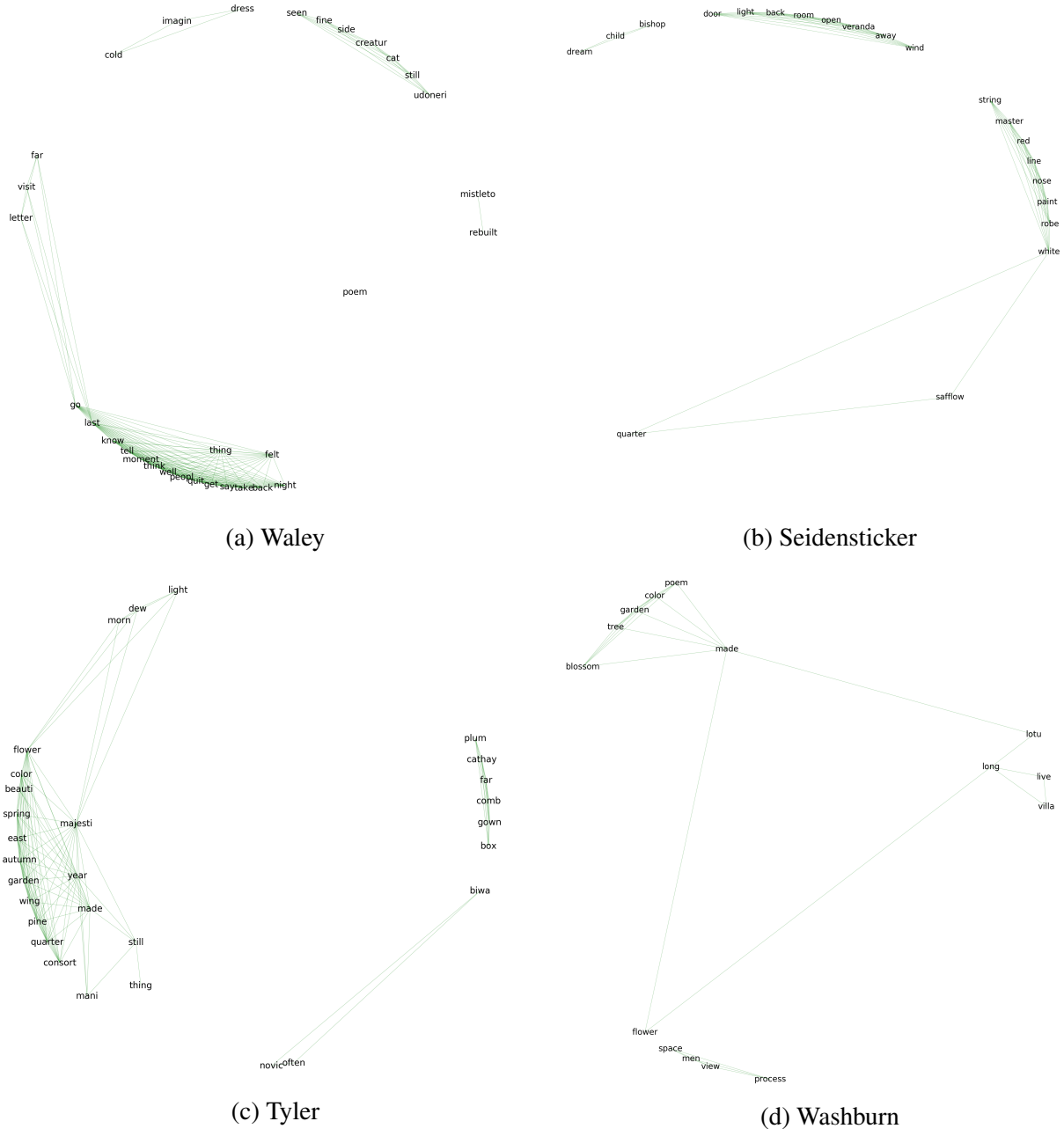
(a) Waley

(b) Seidensticker

(c) Tyler

(d) Washburn

Figure 6: *Genji* Without Women



(a) Waley (b) Seidensticker

(c) Tyler (d) Washburn

Figure 7: *Genji* Without Men

of appearance and service. In Tyler, ceremonial words like ’rever’, ’music’, ’play’, ’majesti’ and ’high’ occur with high degrees, Waley’s has similar trends with ’palace’ and ’prince’, and Seidensticker’s features ’look’, ’governor’, and ’long’. From these observations, we can identify three major categories as bases for interrogating the model. Feeling and thinking: ’feel’, ’think’, ’felt’, ’thought’, ’tear’, ’emot’, ’die’, performance and appearance: ’perform’, ’play’, ’dress’, ’robe’, ’music’, and duty and court: ’palace’, ’ceremoni’, ’attend’, ’rever’, ’crown’.

The most intriguing changes, however, are when we remove individual female words rather than all female words in their entirety. Removal of the word ’mother’ has different effects on each network, reducing Seidensticker’s to a notably more romantic, sentimental, and courtship-heavy structure, with heavy occurrences of the feeling and thinking words. The structure remains intact, with no severed portions of the network. In Waley’s case, removal of the word ’mother’ results in a collapsing of the network, suggesting a much heavier reliance on women as they relate to men. A similar collapse is not observed if we remove more individual female-associated words like ’girl’ or ’woman’.

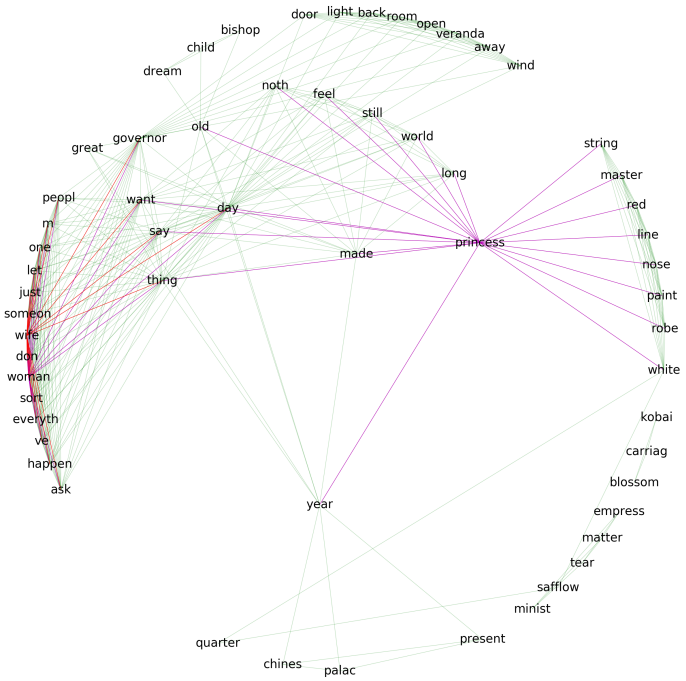


Figure 8: Removing ’Mother’ from Seidensticker

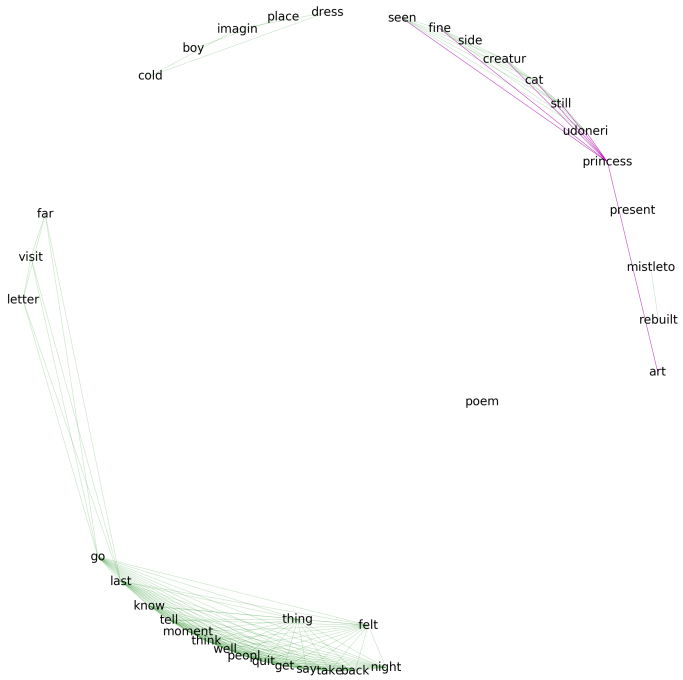


Figure 9: Removing ’Mother’ from Waley

Works Cited

Jockers, M. L..*Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press, 2013. Project MUSE.

Rhody, Lisa M. "Topic Modeling and Figurative Language." *Journal of Digital Humanities*, vol. 2, no. 1, 2012, pp. 19—38.

Henitiuk, Valerie. "Going to Bed with Waley: How Murasaki Shikibu does and does not become world literature." *Comparative Literature Studies*, vol. 45, no. 1, 2008, pp. 40—61., www.jstor.org/stable/25659632.

Meeks, Elijah. "Using Word Clouds for Topic Modeling Results." *Digital Humanities Specialist*. N.p., 15 Aug. 2012. Web. 16 Apr. 2017. ¡https://dhs.stanford.edu/algorithmic-literacy/using-word-clouds-for-topic-modeling-results/¿.

Shikibu, Murasaki, and Dennis C. Washburn. *The Tale of Genji*. New York: W.W. Norton, 2015. Amazon.com. Web.

Shikibu, Murasaki, and Arthur Waley. *The Tale of Genji*. Boston and New York: Houghton Mifflin, 1925. Amazon.com. Web.

Murasaki Shikibu B. "University of Oxford Text Archive." [OTA] *The Tale of Genji* [Electronic Resource]. Trans. Edward G. Seidensticker. N.p., n.d. Web. 16 Apr. 2017. ¡http://ota.ox.ac.uk/desc/2245¿.

Murakami, Janel R. Goodman. "Metonymy in The Tale of Genji: An Analysis of Translation Strategies." *Arizona Working Papers in SLA and Teaching 20* (2013): 55-75.

Moretti, Franco. "Network Theory, Plot Analysis." *Distant Reading*. London: Verso, 2015. N. pag. Print.