Table of Contents for Vol. 2, No. 1 Winter 2012

Topic Modeling and Figurative Language

Lisa M. Rhody

…to have them for an instant in her hands both at once,
the story and its undoing…
from "Self Portrait as Hurry and Delay" [Penelope at her loom]

Located at the center of Jorie Graham's collection *The End of Beauty*, "Self Portrait as Hurray and Delay" crafts a portrait of the artist, poised at a precarious moment in which thought begins to take shape. Like Penelope, Graham entertains the illusion, if only momentarily, of a choice between bringing a creative impulse into form or allowing it to come undone. A weaver of language, Graham subtly, deftly, but unsuccessfully attempts to delay the inevitable moment in poetic creation in which complexity of thought adopts form through language, and so

realized is also reduced. In *The End of Beauty*, the beginning of the creative act signals an inevitable descent into meaning — language's ultimate impulse.

Understanding how topic modeling algorithms handle figurative language means allowing for a similar beautiful failure — not a failure of language, but a necessary inclination toward form that involves a diminishing of language's possible meanings. However, the necessarily reductive methodology of sorting poetic language into relatively stable categories, as topic modeling suggests, yields precisely the kind of results that literary scholars might hope for — models of language that, having taken form, are at the same moment at odds with the laws of their creation.

In the following article, I suggest that topic modeling poetry works, in part, because of its failures. Somewhere between the literary possibility held in a corpus of thousands of English-language poems and the computational rigor of Latent Dirichlet Allocation (LDA), there is an interpretive space that is as vital as the weaving and unraveling at Penelope's loom.

When Michael Witmore refers to texts as "massively addressable at different levels of scale," as he does in his two blog posts in *Debates in the Digital Humanities* (2012), he taps into a similar vein of thought as Jorie Graham. Witmore explains that

> What makes a text a text — its susceptibility to varying levels of address — is a feature of book culture and the flexibility of the textual imagination. We address ourselves to this level, in this work, and think about its relation to some other. (325)

In other words, texts can be approached from a multiplicity of perspectives — as bound entities, pages, chapters, paragraphs, poems, or "works." Textual and literary scholarship requires a willingness to isolate a particular aspect of the text through often abstract or arbitrary constraints, producing what Witmore calls "unities." To a certain extent, textual scholarship implies a double bind: no one can address a text at all of its possible levels simultaneously, and yet, by constraining our understanding of what a text is, we make a caricature of it. Witmore describes "narrowing" our perspective of a text in caricature as "willfully abstract in the sense that, at crucial moments of the analysis, we foreground relations as such — relations that should be united with experience" (329).

The constraints of choosing one textual "unity" correspondingly expands our ability to address a larger scale of texts, revealing patterns and relationships that might otherwise have remained hidden. By locating "figurative language" as an aspect of address for topic modeling, I choose to constrain my consideration of poetic texts and agree to a caricature of poetry that hyper-focuses on its figurative aspects so that we can better understand how topic modeling, a methodology that deals with language at the level of word and document, can be leveraged to identify latent patterns in poetic discourse.

Revising Ekphrasis

Topic modeling with LDA first captured my attention as a possible way to ask discovery-oriented questions about a genre of poetry called ekphrasis — poems written to, for, or about the visual arts. Contemporary critical models of ekphrasis define the genre through the identification of recurring tropes invoked by poets confronted by the differences between linguistic and visual media. Drawing from a longstanding tradition of competition between poets and painters and the verbal and visual arts, our most recognized critical model for ekphrasis turns on the axis of difference, otherness, hostility, and competition. Conventions of ekphrasis include vocalizing the poet's frustrated desire for the still, fixed, and feminized image ("Ode on a Grecian Urn" by John Keats); narrating the pregnant moment of the visual work of art ("Landscape with the Fall of Icarus" by William Carlos Williams); recounting one's visit to a museum as if the reader's guide or teacher ("Musée des Beaux Arts" by

W.H. Auden); describing a figure transfixed on the canvas ("My Last Dutchess" by Robert Browning); or even using the image as a vehicle to travel back through public and personal history ("For the Union Dead" by Robert Lowell). Much like my abbreviated list here, the "canonical" texts used to trace the long-standing tradition of ekphrasis, from Homer's first description of Achilles' shield in the *Illiad* to John Ashbury's "Portrait in a Convex Mirror," have been based until just recently on examples exclusively by men.

LDA, then, offered an attractive alternative for asking questions about the ekphrastic tradition for two reasons. First, as a computational method it allowed me to cast a much wider net. Rather than selecting from just a few poems, LDA allowed me to cast my net as wide as 4,500 poems. Second, both LDA and our existing model of ekphrasis presuppose that latent patterns of language, when discovered, can be used to describe the corpus as a whole. Organizing a corpus of poetry in terms of its participation in recognized conventions of language seemed in keeping with LDA's assumptions that texts are composed of a fixed number of topics, and so I was drawn to the prospect of using LDA to uncover ways poets enter into, disrupt, or perpetuate the ongoing discourses associated with the tropes that typify ekphrasis.

Therefore, the rationale for deploying LDA as a method of discovery and as a means of understanding the contents of large corpora of texts begins with a similar set of assumptions. For example, LDA assumes that text documents in large corpora tend to draw from categories of language that are associated with the subjects of those documents. In an effort to discover the semantic composition of a large collection of text documents, LDA calculates the likelihood that words that refer to similar subjects appear in similar contexts, and then the LDA algorithm groups those words into "topics." LDA, then, presupposes that we can discover the semantic composition of a corpus by grouping into "topics" distributions of words from a set vocabulary that tend to occur together. The process is not unlike the critical assumptions made about ekphrasis — that it draws repeatedly from the same tropes and conventions.

Unpacking the Assumptions of LDA

Following in the vein of Matthew Jockers, Ted Underwood, Scott Weingart, and others who have published gentle introductions to topic modeling for humanists,[1] I want to begin with a short, if potentially reductive, narrative of how LDA generates topics from text corpora. I will return to this example throughout the article to illustrate how highly figurative language texts such as poetry respond to LDA differently than texts that strive for more literal meaning.

Imagine that there is a farmers' market on the other side of town. Many of your neighbors rave about the quality of the produce there, but you would like to know what kinds of produce are available before you decide to drive across town to try it out. One Saturday morning, your neighbors leave for the market with empty baskets and return with full baskets. You assume that your neighbors can only choose from the types of produce available at the farmer's market and that there is a limited variety of produce available. Since it is happens to be late summer in our fictional story, your neighbors select from 10 types of produce that are available at the market: early Gala and Granny Smith apples, butternut squash, Bosc pears, and one neighbor even snatches up the last pint of blueberries. One by one as your neighbors return, you survey the contents of their baskets. Looking into more and more baskets and revising your predictions, you reconsider based on which produce appears together in a basket the most frequently how to reorganize the 10 produce types.

Examining the quantities and varieties of produce in each basket, you could begin to predict not only the range of produce that might have been at the farmers' market but also the relative quantities. Over the course of sampling your neighbors' baskets, you come to the conclusion that the selection of produce at the farmer's market consists of 20% green apples, 20% red apples, 15% pears, 10% winter squash, 10% cantaloupe, 5% corn, 5% beans, 5% tomatoes and 5% assorted other kinds of produce that were different enough from one another

that it makes sense to just call them miscellaneous. As more neighbors arrive, with baskets to examine, you can refine your predictions about what the available selection of produce have been at the market.

In the case of the farmer's market, your approach to predicting the 10 kinds of produce and the available quantities of each based on the contents of your neighbor's baskets is akin to the way LDA algorithms approach texts. LDA assumes that documents are like your neighbors' baskets, and your neighbors are like authors who select from a limited number of available types of words in order to produce documents — in this case poems. Each author chooses to varying degrees how much of each kind of topic they use for each document; however, the number of total available topics, just like the total number of kinds of produce remains constant. While this constraint, the assumption that all the words in a corpus could be derived from a limited set of topics, strikes the human reader as an artificial limitation, it is a necessary constraint in order for LDA to work.

LDA attempts to describe the overall distribution of topics in a collection of texts in the same way that you discovered the types and quantities of produce at the market. The size of the "topics" likewise reflects your estimation of how much of each kind of produce is available. You were able to predict that there were more apples and pears at the market than there were blueberries and tomatoes because across the whole sampling of baskets there were more apples and pears and fewer pints of blueberries.

There is one significant difference, however, between the human topic model example and the algorithm. LDA does not produce names for the topics it discovers or sort words with an understanding of what words *mean*. Imagine that while you are sorting through baskets, you come across an Asian pear. You've never seen an Asian pear before, but the Asian pear was in a basket with a large number of apples and pears. You make note of that, set it in either the apple or pear group temporarily, knowing that you will come back to it after you have gathered more information and continue to sort through baskets. Over the remaining baskets, Asian pears tend to appear in other baskets where there are also other kinds of pears more often than in baskets where there are also apples. As a result, you come to the conclusion that, since Asian pears frequently appear in baskets with other pears, the Asian pear in each future basket should be sorted with the pears. This method of determining how to sort Asian pears reflects the manner in which LDA assigns words to topics, according to the other words that are found in the same document. Although the algorithm cannot account for what words mean, much like your method of discovery about Asian pears, LDA does a surprisingly good job of sorting words based on co-occurrence. Finally, LDA sorts words into topics based on prior knowledge that there are a finite number of topics in the overall corpus — much the same way that you knew to look for 10 types of produce.[2]

Topic models (and LDA is one kind of topic modeling algorithm) are generative, unsupervised methods of discovering latent patterns in large collections of natural language text: generative because topic models produce new data that describe the corpora without altering it; unsupervised because the algorithm uses a form of probability rather than metadata to create the model; and latent patterns because the tests are not looking for top-down structural features but instead use word-by-word calculations to discover trends in language. David Blei, credited with developing LDA and probabilistic topic modeling methods, describes topic models the following way:

> Topic models have been developed with information engineering applications in mind. As a statistical model, however, topic models should be able to tell us something, or help us form a hypothesis, about the data. What can we *learn* about the language (and other data) based on the topic model posterior? (Blei "Introduction" 84)

Blei stages topic modeling as an *ex post facto* method for challenging our assumptions about natural language data. In other words, once a collection has been created, LDA can test our assumptions about what topics are discoverable.

What drew me to LDA as a tool for discovering latent patterns of language use in ekphrastic poetry was that it seemed particularly well-suited to identifying the tropes of ekphrastic discourse. One could reasonably expect that since the language of stillness, breathlessness, desire, and competition are commonly found in ekphrastic poetry, that LDA might be able to locate ekphrastic poems within a much larger corpus — in this case 4,500 poems. I wondered, could topic models detect gendered language, tropes, or the language of stillness in ways that "we can *learn*" about the genre more broadly? This is the question that began *Revising Ekphrasis*, a digital topic modeling and corpus discovery project I developed that uses digital and computational tools to explore ekphrastic and non-ekphrastic poetry.

The topic model represented in this article is one of several from the *Revising Ekphrasis* project. I've chosen this particular model to focus on for two reasons. It was the first model in the project to produce results that prompted a reconsideration of the tropes and conventions of ekphrasis. Secondly, it illustrates how figurative language resists thematic topic assignments and by doing so, effectively increases the attractiveness of topic modeling as a methodological tool for literary analysis of poetic texts. Few questions will find "answers" here. Instead the hope is to uncover new methods for addressing enduring humanities questions that we might fruitfully ask about figurative language with LDA.

LDA Topics and Poetry

A form of text mining developed in response to the growing challenge of managing, organizing, and navigating large, digitized document archives, topic modeling was developed with primarily non-fiction corpora in mind. One of the most notable, early uses of LDA by Blei explores a digitized archive of the journal *Science*. Other exemplary topic modeling projects have used Wikipedia, NIH grants, JSTOR, and an archive of Classics journals.[3] As literary scholars well know, however, poems exercise language in ways purposefully inverse to other forms of writing, such as journal articles, encyclopedia entries, textbooks, and newspaper articles. Consequently, it is reasonable to predict that there will be differences between topics created by LDA models of poetry and models of non-fiction texts. In terms of the non-figurative language found in topic models of the journal *Science*, Blei explains that topics detect *thematic* trends across texts:

> We formally define a topic to be a distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. (Blei "Introduction" 78)

Presented as a method of discovery and description, computer scientists see topics as revealing latent thematic trends that pervade large and otherwise unstructured text corpora, and with respect to the data used to create the topic model, this conclusion makes sense and works well.

Since topic modeling was designed to be used with texts that employ as little figurative language as possible, the expectation that words with similar meanings will be found in the same document as other words with related meanings makes sense. This is not the case, however, in a genre like poetry, where the use of highly figurative speech actually increases the scope of the language one might expect to see in a document. For example, literary devices such as metaphor or simile compare two objects, experiences, or feelings that are completely unalike, and in doing so isolates and heightens our awareness of what makes them similar. Poetic texts are more likely to contain purposefully-figurative language; therefore, the first step in understanding how figurative language responds to LDA is to consider what changes occur between the topic assignments in a journal article from *Science* in direct contrast to the same process for a poetic text — in this case, Anne Sexton's "The Starry Night."

In order to compare how LDA creates topics in non-figurative texts (*Science)* versus how topics are generated from a corpus of poetry, I begin with an overview of how Blei's model of 100 topics across 17,000 *Science* articles are created. Next, I create a parallel example using Anne Sexton's poem "The Starry Night" from a 60

topic model of 4,500 poems from the *Revising Ekphrasis* dataset, pointing out how topic models estimate topic proportions in the document and how topic keyword distributions in poetry are not "thematic" in the way that topic models of non-fiction documents are.

In "Probabilistic Topic Models," Blei uses two illustrations to explain how topic modeling of a large, digitized collection of *Science* works. The first illustration depicts an excerpt from one article within the collection titled "Seeking Life's Bare (Genetic) Necessities" and demonstrates the relationship between topics and keyword distributions. His first illustration (Figure 1) uses the colors yellow, pink, green, and blue to represent four of the topics that the model predicts exist in the dataset. Recalling my earlier example of the farmers' market, the pink, blue, and yellow topics are like the types of produce at the market. On the far right hand side of Figure 1 is a bar graph that represents the proportions of the yellow, pink, and blue topics the model predicts are in the document (an article in this case). The largest topic in the document is yellow followed by pink then blue. The lines from the bar graph on the far right point to the places in the text where words that are associated with the yellow, pink, and blue topics can be found in the document. Essentially, the histogram in Figure 1 is showing the equivalent in the farmers' market example of there being more apples than pears or grapes in a single basket. On the far left hand side are the first three words of the topic keyword distribution. Those represent the individual produce items in each produce type that could be found in the places in the text that are highlighted in yellow, pink, and blue.

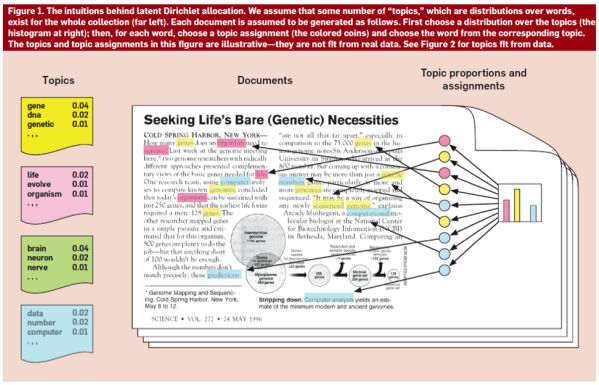

Figure 1: Illustrative example of *Science* topic model (Blei "Introduction" 78)

The graphic in Figure 1 helps to identify how the topic proportions (like the number of apples in a basket of produce from the market) correlate to individual words in the document (highlighted above in yellow, pink, and blue), which then comprise the "topic" keyword distributions that are displayed at the far left as a partial list of keywords.[4]

Figure 1 is an illustrative example, meaning the document and topic assignments in the graphic are not actually derived from a specific model; however, in a second graphic, Blei continues to explain the how "Seeking Life's Bare (Genetic) Necessities" appears within a 100 topic model of 17,000 *Science* articles. In Figure 2, Blei represents the probability of each topic using a histogram (bar graph) that demonstrates the relationship between the topics 0-99 (along the horizontal axis) and the probability (as a decimal along the vertical axis) that the topic is found in "Seeking Life's Bare (Genetic) Necessities." Some topics have higher probabilities of appearing in the document than others, as represented by the taller bars in the graph. On the right side of the graphic, the topic keyword distributions are listed vertically in columns. At the top of each column is a bolded word surrounded by quotation marks that serves as a label created by Blei to describe the words in the topic and demonstrating Blei's

rationale for claiming that topics are thematic. For example, the topic labeled "Genetics" is predicted by LDA to be the largest topic in the document in much the same way that in the farmer's market analogy you could determine that the largest produce type in a single basket was from the topic "apples." In that light, the model's prediction about "Seeking Life's Bare (Genetic) Necessities" makes sense. We would normally expect the words human, genome, dna, genetic to be found in articles about "genetic necessities." By glancing over the words in the topic keyword distributions, we gather together a sense of what the article might be about.

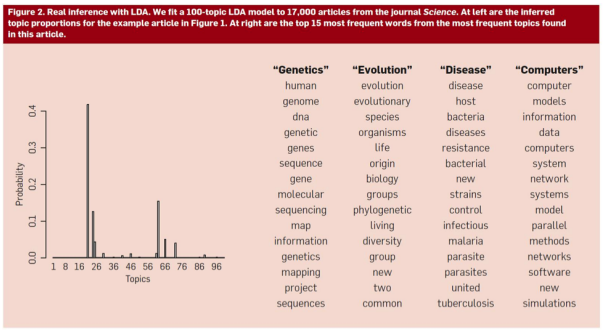

Figure 2: Topic keywords for a single document in *Science* and the proportion of the document described by each topic.

Surveying Blei's list of key terms in each topic in Figure 2 clarifies the way in which models predict thematic trends in large text corpora. The sense that each of the words in each of the columns belongs together makes a compelling case for LDA's ability to use Dirichlet allocation to sort large collections of documents into topical categories. Affixing the term "latent" to the statistical model (latent Dirichlet allocation), as Blei explains, foregrounds the expectation that topic modeling is meant to discover hidden patterns within the large collection of texts. It would take even the most proficient human reader an extraordinary period of time to read 17,000 articles from *Science*. Therefore, while we know through disciplinary familiarity and deduction that the topics in Figure 2 are likely topics to be found throughout the journal's publication, we wouldn't be able to detect or retain those patterns through human reading. Blei, therefore, concludes that probabilistic topic modeling "provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text" ("Introduction" 82).

Unsurprisingly, humanists interested in sorting, sifting, and organizing large collections of text, managing large document archives, and creating better browsing options for digital libraries find LDA's potential exciting and promising. Furthermore, humanists interested in uncovering the "latent patterns" in large datasets are likewise enthused by the algorithm's potential for exploratory studies. Most notably, Robert Nelson's project *Mining the Dispatch* employs LDA to uncover hidden patterns within the archives of the *Richmond Daily Dispatch* just before, during, and after the Civil War. Nelson's LDA analysis uses the topic distributions over thousands of *Dispatch* articles over the course of the war to track relationships between increases in military draft and fatalities and the patriotic rhetoric. Even more impressively, Nelson's utilization of LDA is more than a descriptive endeavor because he moves from identifying topic distributions to engaging humanities concerns such as shifts in the rhetoric of nationalism in the Confederate South during the Civil War in relationship to changes in casualty rates and calls for enlistment.[5] Nelson's work in this area represents one of the most ambitious and successful projects to date in the humanities that uses probabilistic topic modeling. *Mining the Dispatch* is the first to broach the territory of figurative language and LDA in its analysis of patriotic discourse in Civil War Confederate newspapers. In Nelson's project, poetry is combined with opinion articles and political and agricultural reports, and the composition of the dataset seemingly allows the poetic texts to map well with its prose counterparts.

However, topic models of purely figurative language texts like poetry do not produce topics with the same *thematic* clarity as those in Blei's topic model of *Science* or even Nelson's model of the *Richmond Daily Dispatch*. The literary scholar has good reason to be skeptical about the results of LDA analysis when the dataset to be explored includes primarily, if not exclusively, poetic texts. Given our disparate expectations for how language should operate in poetry as opposed to non-fiction, should the same standards for evaluating topic models of non-figurative language texts guide the principles we use to evaluate the accuracy of topic models of figurative language collections? How would they differ?

Evaluating Topic Models of Figurative Language

As Ian H. Witten, Eibe Frank, and Mark A. Hall remind us in *Data Mining: Practical Machine Learning Tools and Techniques,* the guiding factors for text mining generally and topic modeling specifically are to generate *actionable* and *comprehensible* results (9.5).

> **Actionable**: Results should be consistent and reproducible, which means that the model could also be used to make predictions about new data added to the dataset. Of course, whether or not results are indeed actionable depends to a large extent on the ability to find a fair and measurable degree of success. Actionable results require that researchers are clear about their *a priori* assumptions and the composition of the dataset and the predicted degree to which the results might be found reliable.
>
> **Comprehensible**: For the results of text mining to be useful, humans need to be able to read, to understand, and to interpret them. Frequently, in topic modeling comprehensible results are understood to be thematic or semantically meaningful. In other words, when reading key word distributions, it is usually obvious that there is a thematic array that humans can read and interpret sensibly. For example, in Blei's keyword distributions the terms "evolution, evolutionary, species, organisms, life, origin" lead to a comprehensible thematic topic: evolution.

Herein lies the rub for texts as highly figurative, purposefully ambiguous, and semantically rich as poetry. Returning once again to Blei's article, he writes: "The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents," which he clarifies further in a footnote:

> Indeed calling these models "topic models" is retrospective — the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA. (Blei "Introduction" 79)

The topics from *Science* read as comprehensible, cohesive topics because the texts from which they were derived aim to use language that identifies very literally with its subject. The algorithm, however, does not know the difference between figurative and non-figurative uses of language. So the process LDA employs does not change: topics remain a distribution of words over a fixed vocabulary, such that topics are formed only by those words included in the dataset and in the statistical distribution of those words across the entire set. Therefore, *comprehensible* results, in the case of *Science,* seems a reasonable determiner as to whether or not a model is also *actionable*.

What, if anything, changes if we work through a parallel example of how a topic model "reads" Anne Sexton's "The Starry Night"? The model used for this example used 4,500 poems from the *Revising Ekphrasis* dataset to generate 60 topics. When the collection of poems was prepared for the experiment, words that hold a relatively small amount of semantic weight, but are numerous enough to skew the model's results, such as articles,

frequently used pronouns, conjunctions, prepositions, and pronouns were removed. In the example below, the words removed before the topic model was run have been struck out.

Returning to the farmer's market example from earlier in this article, "The Starry Night" is an example of what one neighbor's basket of produce (poem/document) might look like. The basket's contents are distributed much like the produce in the neighbors' baskets. 29% of the produce (words) would be like apples (Topic 32), 12% of the produce would be corn (Topic 2), and 9% of the produce would be like grapes (Topic 54).[6] All in all, 50% of the basket (poem/document) can be accounted for by three produce types (topics).[7] For simplicity's sake, I have ignored the smaller topics and will focus just on the top three topics found in the document. In order to simulate to some degree the way in which the topic model "reads" the poem, I have crossed out words that would be removed by the stoplist, and highlighted in green (Topic 32), yellow (Topic 2), and blue (Topic 54).

In Table 1, which directly follows the poem, there are three columns that list the topics from which "The Starry Night" is predicted by the LDA to draw most heavily. In each column of the table, the number of the topic is listed at the top next to the probable proportion of the document that uses words from this topic. The fifteen words below each Topic number represents a sampling of the word distribution that makes up the whole topic. For example, in the farmer's market example the topic with the largest percentage would be "apples." Under the "apples" topic, we might find Macintosh, Fuji, Honeycrisp, and Gala, all words associated with apples. For the purpose of making the assignment of words from the poem to the topic keyword distributions clear, each topic has been assigned a color (green/32, yellow/2, blue/54).[8]

~~The~~ Starry Night

~~That does not~~ keep ~~me from having a~~ terrible need ~~of—shall I say the word—~~religion. ~~Then~~ I go ~~out at~~ night ~~to~~ paint ~~the~~ stars. Vincent Van Gogh ~~in a~~ letter ~~to his~~ brother

~~The~~ town ~~does not~~ exist
~~except where one~~ black-haired tree slips
~~up like a~~ drowned woman ~~into the~~ hot sky.
~~The~~ town ~~is~~ silent. ~~The~~ night boils ~~with~~ eleven stars.
~~Oh~~ starry starry night! ~~This is how~~
~~I want to~~ die.

~~It~~ moves. ~~They are all~~ alive.
~~Even the~~ moon bulges ~~in its~~ orange irons
~~to~~ push children, ~~like a~~ god, ~~from its~~ eye.
~~The old~~ unseen serpent swallows ~~up the~~ stars.
~~Oh~~ starry starry night! ~~This is how~~
~~I want to~~ die.

~~into that~~ rushing beast ~~of the~~ night,
sucked ~~up by that~~ great dragon, ~~to~~ split
~~from my~~ life ~~with no~~ flag,
~~no~~ belly,
~~no~~ cry.

Figure 3: "The Starry Night" by Anne Sexton. Text with a strike through it has been removed as a stopword during preprocessing. Text highlighted in green can be found in Topic 32. Text highlighted in yellow can be found in Topic 2. Text highlighted in blue can be found in Topic 54.

Table 1: Keyword distributions generated by a 60 topic model of 4500 poems (Note: Keywords in this table are representative of the entire model, not just words from "The Starry Night."

Topic 32	Topic 2	Topic 54

night	death	tree
light	life	green
moon	heart	summer
stars	dead	flowers
day	long	grass
dark	world	trees
sun	blood	flower
sleep	earth	spring
sky	man	leaves
wind	soul	sun
time	men	fruit
eyes	face	garden
star	day	winter
darkness	pain	leaf
bright	die	apple

Topic 32 and 54 appear similar to the coherent, thematic topics in the topic model of *Science*. Topic 32 includes words that could fall under the rubric of "night," and the words in Topic 54 could be described as the "natural world." We might be tempted based on this first read to assign the topic labels "night" and "natural world" in the same way that Blei labels topics from *Science* as "genetic" and "evolution"; however, as I will discuss further on, those labels and the assumption that the topics are "thematic" in the same way as Blei's would be incorrect. For example, the night and natural world of "The Starry Night" are actually painted representations of those concepts, and consequently, it would be misleading to say that the poem is, strictly speaking, about night and the natural world *in the same way* that the article from *Science* is about genetics and evolution.

Topic 2, on the other hand, does not have the same unambiguous comprehensibility that 32 and 54 do: the words in Topic 2 are more loosely connected. It would be tempting to read the topic as having to do with death, but we would do that because our reading of "The Starry Night" predisposes us to consider it that way. There are "intruder" words in this category. By looking solely at the words in the list and not taking into consideration "The Starry Night," words such as long, world, and day are not necessarily words we might classify as "death" words in the strictest sense.

In fact, topic intrusion is one way in which computer scientists have begun to develop a method for evaluating and interpreting topic models. In "Reading Tea Leaves: How Humans Interpret Topic Models," (pdf) Jonathan Chang, Jorden Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei suggest methods for measuring the "interpretability of a topic model" (2). The authors present two human evaluation tests meant to discern the accuracy of models by using the keyword distributions (like the individual items from the farmers' market), and the topic to document probabilities (the proportion of kinds of apples compared to how many fruit are in each basket) — called word intrusion and topic intrusion tests respectively. Word intrusion tests involve selecting the first eight or so words from each topic and adding one word to each list for a total of nine words. Human subjects (generally disciplinary experts) were then asked to determine which word in each group did not belong. Chang, et al. discovered that with relative high success, human readers could discern a thematic connection between terms to reliably distinguish the single out-of-place term. As a result, the authors suggest that word intrusion tests measure "how well the inferred topics match human concepts" (6).

On the other hand, topic intrusion tests presented human subjects with topic labels (like apples, pears, and corn are labels for the "types of produce" that might be at the farmer's market); the words most likely to be associated with each topic (such as Macintosh, Gala, Fuji, and Honeycrisp), and the top documents associated with each

topic (basket number 1, basket number 2, basket number 3, for example). Then, one document (a basket unlike any of the others) that does not belong in the group, the "intrusion," is then added to the set, and human subjects were then asked to identify which document did not belong, which, again, they could do with reasonable accuracy.

For the purposes of modeling poetry data, word intrusion would not be as effective a method for determining a model's accuracy at categorizing documents or detecting latent patterns unless the specific changes that happen to the nature of topic distributions for poetic corpora are adjusted for. "Intruders" as individual words does not work for LDA topics of poetry because poems purposefully access and repurpose language in unexpected ways. In other words, topics from the models in my project were not easily interpreted by keywords alone, and yet the results are still useful.

Interpreting Models of Figurative Language Texts

Topic models of poetry do have a form of comprehensibility, but our understanding of coherence between topic keywords needs to be slightly different in models of poetry than in models of non-fiction texts. My research confirms, to a degree, Ted Underwood's suspicion that topics in literary studies are better understood as a representation of "discourse" (language as it is used and as it participates in recognized social forms) rather than a thematic string of coherent terms.[9] However, because the topic model I describe here has been "chunked" at the level of individual poems, the matter of how we interpret a model and how we use it as a vehicle for discovery is different from how Underwood deploys the term at the beginning of his interpretive process. My use of the term "discourse" drives my attention back to close readings of individual poems searching for similarities and differences between poems predicted to contain higher proportions of the same topic.

Topic models of poetry do not reflect the anecdotal evidence that LDA frequently leads to semantically meaningful word distributions. Instead, topic models of the *Revising Ekphrasis* dataset created four consistently recurring *types* of topics. Moreover, recognizing the following four types of topics coupled with close reading of samplings of documents containing each "topic," which allows a literary scholar to see coherence in topics as forms of discourses, worked much better for determining whether or not the results of the model were actionable and comprehensible. When viewed as forms of discourse, topics can be re-considered in light of whether or not close readings show that individual documents are entering into a form of discourse for a thematic purpose.

LDA topics from a model of the poetic documents in the *Revising Ekphrasis* dataset return one of four *types* of topic, which I define as follows:

OCR[10] ***and other language or dialect distinctive features***[11] — These topics represent, for example, errors that occur in the optical character recognition scanning process used when turning print documents into digitizing texts, for example substituting "com" for "corn." The most common OCR errors have been filtered out through a preprocessing technique that searches for such errors and fixes them; however, machines aren't perfect and some of these features remain in the final dataset. Their presence may sort out as if they were features of another language. More commonly in this dataset, however, one or two topics form around an approximate 1% of the data that includes foreign language terms or the original form of a poem before its English language translation. The following two topic examples found in the same topic model as "The Starry Night" demonstrate how the model clusters these:

> Topic 4: de la el en green verde con los mi se del poem n lo os poema yo oo ya sobre
>
> Topic 30: de miss ain jump dat ah dey ter yo slim scarlett hunh git back tu stan fu huh barbie den

Similarly, topics can also be created by grouping together distinctive dialects and languages other than English. We will not be considering these topics in detail other than to point out that they exist.

Large "chunk" topics — Longer or extended poems that outsize the majority of other documents in the subset pull one or more topics toward language specific to that particular poem. For example, the keyword distribution for Topic 12 includes terms such as: bongy, yonghy, bo, lady, jug, order, jones and jumblies. These are words that are repeated frequently in the extended poem "The Courtship of the Yonghy-Bonghy-Bo" by Edward Lear and demonstrate how one poem with high levels of repetition can pull a topic away from the rest of the corpus, along with other poems with high frequency repetitions of particular phrases. In the case of Topic 12, the poems included in the topic and shown in Table 2 tend to be longer and to include greater incidence of repetition. It is possible that these poems share thematic affinities, but the strength of those affinities have more to do with linguistic structure than meaning. In Table 2, the documents with the highest probabilities of drawing a large proportion of their words from Topic 12 are listed in descending order. Under the "Topic 12" label are the probable proportions for each document expressed in decimals. In the second column are the corresponding poem titles.[12]

Table 2: Titles of poems in the Revising Ekphrasis *dataset with the highest probable proportion of Topic 12, listed in descending order. In the list of poems, those available on the American Academy of Poets website (www.poets.org) can be reached by clicking the link on the poem's title.*

Topic 12	Poem Title
0.680665	The Courtship of the Yonghy-Bonghy-Bo
0.590501	Choose Life
0.504747	Zero Star Hotel [At the Smith and Jones]
0.501921	The Midnight [For here we are here]
0.47986	Earthmover
0.462247	Invitation to the Voyage
0.412626	Mr. Macklin's Jack O'Lantern
0.358385	The Steel Rippers
0.333965	The Cruel Mother

0.276595	Vacant Lot with Pokeweed
0.274312	Lullaby of an Infant Chief
0.253223	The Jumblies
0.250493	American Sonnet (35)
0.230571	Rückenfigur
0.221246	Two Poems
0.217995	The Lady of Shalott
0.2177	Mr. Smith
0.209471	The Assignation
0.191892	Ulalume
0.179114	I Too Was Loved by Daphne

Semantically *evident topics* — Some topics do appear just as one might expect them to in the 100-topic distribution of *Science* in Blei's paper. Topics 32 and 54, as illustrated above in Anne Sexton's "The Starry Night," exemplify how LDA groups terms in ways that appear upon first blush to be thematic as well. As I mentioned earlier, though, the illusion of thematic comprehensibility obscures what is actually being captured by the topic model. The way in which we interpret semantically evident topics like 32 and 54 must be different from the semantically coherent topics of non-figurative language texts. It is more accurate to say that Topics 32 and 54 participate in discourses surrounding "night" and "natural landscapes" in Anne Sexton's "The Starry Night."

As Elizabeth Bergmann Loizeaux points out in *Twentieth-Century Poetry and the Visual Arts*, Sexton's poem enters into an ongoing conversation with other confessional poets about madness and artistic genius by engaging in language that refocuses collective attention on a widely-recognized work of art with a recognized connection

to another artist suffering from mental duress.[13] She enters into that discourse through the other surrounding discourses that include night and natural landscape. It would still be incorrect to say that 29% of the document is "about" night, when what Sexton describes is a *painting* of a night sky and natural landscape. As literary scholars, we understand that Sexton's use of the tumultuous night sky depicted by Vincent Van Gogh provides a conceit for the more significant thematic exploration of two artists' struggle with mental illness.

Therefore, it is important not to be seduced by the seeming transparency of semantically evident topics. Even though the topics appear to have a semantic relationship with the poems because they appear so comprehensible, it is important to remember that semantically evident topics form around a *manner* of speech that reflects quite powerfully the definition of discourse described by Bakhtin: "between the word and its object, between the word and the speaking subject, there exists an elastic environment of other, alien words about the same object" (293). The significant questions to ask regarding such topics when interpreting LDA topic models have more to do with what we learn about the relationships between the ways in which poems participate in the discourses that the topic model identifies. Word intrusion tests (the kind suggested by Chang, et. al. as a measurement of a model's accuracy) may still work with semantically evident topics because semantically evident topics mirror the thematic comprehensibility of topics from models of non-figurative language; however, there are naturally occurring word intrusions that may not affect the efficacy of the topic distributions, and these would require deeper human interpretation before just throwing them out.

Semantically *opaque topics* — Some topics, such as Topic 2 in "The Starry Night," appear at first to have little comprehensibility. Unlike semantically evident topics, they are difficult to synthesize into the single phrases simply by scanning the keywords associated with the topic. Semantically opaque topics would not pass the intrusion tests suggested by Chang, et. al. because even a disciplinary expert might have trouble identifying the "intruder" word as an outlier. Determining a pithy label for a topic with the keywords, "death, life, heart, dead, long, world, blood, earth…" is virtually impossible *until* you return to the data, read the poems most closely associated with the topic, and infer the commonalities among them.

In Table 3, I list the poems the model predicts contain the highest amount of Topic 2 in them along with the probable proportion of the document that draws from Topic 2 (The amount of each basket the model predicts can be described as "apples," for instance).

Table 3: Titles of the 15 poems with the highest predicted proportions of Topic 2 in them and their corresponding topic distributions. If the poem is available through the American Academy of Poets (www.poets.org), you can read it by clicking on the link from the poem's title.

Topic 2	Poem Title
0.535248643	When to the sessions of sweet silent thought (Sonnet 30)
0.533343438	By ways remote and distant waters sped (101)
0.517398877	A Psalm of Life
0.481152152	We Wear the Mask
0.477938906	The times are nightfall, look, their light grows less
0.472091675	The Slave's Complaint
0.451175606	The Guitar
0.447100571	Tears in Sleep
0.446314271	The Man with the Hoe

0.437962153	A Short Testament
0.433767746	Beyond the Years
0.433152279	Dead Fires
0.429638773	O Little Root of a Dream
0.427326132	Bangladesh II
0.425835136	Vitae Summa Brevis Spem Nos Vetat Incohare Longam

Skimming the top fifteen poems associated with Topic 2 would confirm our assumption that the model has grouped together kinds of poetic language used to discuss death. Topic 2 is interesting for a number of reasons, not the least of which is that even though Paul Laurence Dunbar's "We Wear the Mask" never once mentions the word "death," the discourse Dunbar draws from to describe the erasure of identity and the shackles of racial injustice are identified by the model as drawing heavily from language associated with death, loss, and internal turmoil — language which "The Starry Night" indisputably also draws from.

To say that Topic 2 is *about* "death, loss, and internal turmoil" is overly simplistic and does not reflect the range of attitudes toward loss and death that are present throughout the poems associated with this topic; however, to say that Topic 2 draws from the language of elegy would be more accurate. Identifying that Dunbar's "We Wear the Mask" and "Beyond the Years" draw from discourses associated with elegy supports recent scholarship by Marcellus Blout in his 2007 essay titled, "Paul Lawrence Dunbar and the African American Elegy:"

> I am using a set of terms that point to how I see Dunbar as initiating a *tradition* of African American elegies. I should underscore here that I am not arguing that the African American practice of the elegy is necessarily distinctive from other traditions of the elegy. But I want to suggest that such practice is continuous. Dunbar's poems of the 1890s point us directly to more recent elegies written by African Americans in the latter part of the twentieth century. (241)

By identifying Dunbar's poems in a topic of elegiac language, the topic model supports Blout's claims that Dunbar's poems participate in elegiac discourse as a means of identity formation for African Americans at the turn of the twentieth century. What the topic model (and the close reading prompted by the topics produced by the model) might also help identify is whether or not other poems by contemporary African American poets similarly draw from Topic 2, further supporting Blout's claim that Dunbar "initiates a tradition."

In fact, Dunbar is not the only African American poet included in the list of documents that draw heavily from Topic 2. "The Slave's Complaint" by George Moses Horton (1797-1884) is also included. "The Slave's Complaint" moves through the three stages one might expect to find in an elegiac poem — from lamentation to praise to possible consolation. Could Horton, a poet and a slave, whose poems were written down by school children and printed under the title *The Hope of Liberty* in 1829 have been an influential part of Dunbar's inclination toward the elegiac? It would take a combination of more topic modeling tests and more traditional historical and archival research to answer that question; however, these are the questions we have been hoping topic modeling might help produce.

In other words, opaque topics such as Topic 2 in models that have mixed results prompt the kinds of questions we are looking for as humanists. What this small discovery shows is that topic modeling as a methodology, particularly in the case of highly-figurative language texts like poetry, can help us to get to new questions and discoveries — not because topic modeling works perfectly, but because poetry causes it to fail in ways that are potentially productive for literary scholars.

Just as semantically evident topics require interpretation, determining the coherence of a semantically opaque topic requires closer reading of the other documents with high proportions of the same topic in order to check whether or not the poems are drawing from similar discourses, even if those same poems have different *thematic* concerns. While semantically evident topics gravitate toward recurring images, metaphors, and particular literary devices, semantically opaque topics often emphasize tone. Words like "death, life, heart, dead, long, world" out of context tell us nothing about an author's attitude or thematic relationships between poems, but when a disciplinary expert scales down into close readings of the compressed language of the poems themselves, one finds that there are rich deposits of hermeneutic possibility available there.

Searching for thematic coherence in topics formed from poetic corpora would prove disappointing since topic keyword distributions in a thematic light appear at first glance to be riddled with "intrusions." However, by understanding topics as forms of discourse that must be accompanied by close readings of poems in each topic, researchers can make use of a powerful tool with which to explore latent patterns in poetic texts. For poetry data in particular and literary texts in general, close reading and contextual understanding work together, like the weaving and unraveling of Penelope at her loom, in order to identify relations between texts by shuttling between computational de-familiarization and scholarly experience.[14]

List of Works Cited

Blei, David. "Probabilistic Topic Models." *Communications of the ACM* 55.4 (2012): 77–84. Print.

Chang, Jonathan et al. "Reading Tea Laves: How Humans Interpret Topic Models." *Neural Information Processing Systems (NIPS)*. 2009. Web. 3 Oct. 2012.

Graham, Jorie. *The End of Beauty*. First Edition. Hopewell, NJ: Ecco, 1999. Print.

Graham, Shawn, Scott Weingart, and Ian Milligan. "Getting Started with Topic Modeling and MALLET." *The Programming Historian 2*. Web. 21 Mar. 2013.

Heffernan, James A. W. *Museum of Words: The Poetics of Ekphrasis from Homer to Ashbery*. Chicago: University Of Chicago Press, 2004. Print.

Jockers, Matthew. "The LDA Buffet Is Now Open; or, Latent Dirichlet Allocation for English Majors." *Matthew L. Jockers* 29 Sept. 2011. Web. 29 Oct. 2012.

Loizeaux, Elizabeth Bergmann. *Twentieth-Century Poetry and the Visual Arts*. 1st ed. Cambridge, UK; New York: Cambridge University Press, 2008. Print.

Weingart, Scott. "Topic Modeling for Humanists: A Guided Tour." *the scottbot irregular*. 25 July 2012. Web. 21 Mar. 2013.

Witmore, Michael. "Text: A Massively Addressable Object." *Debates in the Digital Humanities*. Minneapolis, MN and London: University of Minnesota Press, 2012. 324–327. Print.

—. "The Ancestral Text." *Debates in the Digital Humanities*. Minneapolis, MN and London: University of Minnesota Press, 2012. 328–331. Print.

Witten, Ian H, Eibe Frank, and Mark Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Third Edition. New York: Elsevier Science, 2011. Web. 25 Mar. 2013.

[1]For other gentle introductions to LDA for humanists, see Matthew Jockers's blog post "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors" or Scott Weingart's blog post "Topic Modeling for Humanists: A Guided Tour" or Shawn Graham, Scott Weingart, and Ian Milligan's "Getting Started with Topic Modeling and Mallet."

[2]The process of determining the number of topics to tell the model to use is not, as of yet, a standardized procedure. The measure for the "right" topic number is often derived through trial and error. After starting with one number (usually between 40 and 60) one determines how "actionable" and "coherent" the topics that the model produces are, adjusting up and down in subsequent iterations until there is agreement that the best model has been produced.

[3]For more information on how LDA has been used by humanists to detect changing attitudes toward patriotism and nationalism, see: Nelson, Robert K. *Mining the Dispatch*.

[4]In the farmers' market example mentioned earlier in this article, each topic (kinds of produce) is composed of the words (Gala apple, Bosc pear, yellow squash, etc.) in the document (basket). Topic keyword distributions are a list of the words likely to be from a particular topic, in order from most likely to least likely. For humans interpreting topic models, key word distributions are often where the process begins.

[5]For more information on how LDA has been used by humanists to detect changing attitudes toward patriotism and nationalism, see: Nelson, Robert K. *Mining the Dispatch*.

[6]The words "poem" and "document" throughout the remainder of this article are used interchangeably because the dataset consists of individual poems saved as individual plain text documents that include only the title and body of individual poems.

[7]The sum of the three top document probabilities: (29+12+9=50)

[8] Again, to be clear, the keywords in each topic are derived from all the documents in the set of 4,500 that the LDA considers to be part of the topic, so there will be more words in the key word distributions than there are in "The Starry Night." The model assumes that words in the key word distribution are often found in the context of other words also listed in the key word distribution.

[9]I qualify this statement out of recognition that the document types Underwood is modeling are volumes as opposed to individual poems, which may have effects on the degree of reliability with which one can make the comparison. For more on conversations between Ted Underwood and I regarding topics as forms of discourse, see Underwood, Ted. "What Kinds of 'topics' Does Topic Modeling Actually Produce?" and Rhody, Lisa. "Chunks, Topics, and Themes in LDA."

[10]OCR – Optical Character Recognition software visually changes scanned print pages into digitized text.

[11]Topic modeling is frequently used to help discover information in a variety of languages. I choose "other" rather than "foreign" here, since not all "other" languages would be for all researchers "foreign" ones.

[12]When the model outputs the probable proportions for each poem, it expresses that proportion in a decimal. When possible in my discussion of a topic, I convert the decimal to a percentage because that expression of proportion seems more appropriate and avoids statements such as "Rukenfigur" is predicted to contain .23 of Topic 12; however, when I list document probabilities as they have been produced from the model, those same numbers are expressed as decimals.

[13]For more on the ekphrastic conversation between Anne Sexton and W. D. Snodgrass regarding "The Starry Night," see Loizeaux, Elizabeth Bergmann. *Twentieth-Century Poetry and the Visual Arts*.

[14]The author would like to thank the Maryland Institute for Technology in the Humanities, especially Travis Brown, Jennifer Guiliano, and Trevor Muñoz, for the support she received while performing the research that led to this paper.

About Lisa M. Rhody

Lisa Marie Rhody received her Ph.D. in English language and literature from the University of Maryland, where her research was supported by a Maryland Institute for Technology in the Humanities (MITH) Winnemore Dissertation Fellowship. Her research combines advanced computational analysis with traditional literary methods to explore 20th-century poetry and American literature, intersections between visual and verbal media, and women's literature. Currently, she is the project manager for WebWise 2013 at the Roy Rosenzweig Center for History and New Media (RRCHNM). She maintains a digital presence at LisaRhody.com and can be followed on Twitter at @lmrhody.