Chapter 4: Collecting Ekphrasis: Building a Digital Collection of Modern Verse to Study Ekphrasis

By Lisa Marie Antonille Rhody, December 2012

Introduction

Whereas chapters 2 and 3 consider the local, internal, and historically-situated ekphrastic poem as a network of discourses between images and words, poets and visual artists, speakers and readers, and a work of art within its social context, the following two chapters explore the advantages of reading "at a distance," to use Franco Moretti's oft-cited phrase, and within a larger-scale context of thousands of other poems to invigorate and broaden our understanding of ekphrasis—its tradition, tropes, and canon (*Conjectures* 56). In this chapter, I discuss the process of assembling a digital corpus of approximately 4,500 poems, which in the next chapter will become the dataset used for two topic modeling experiments.[63] Because the composition of a dataset determines the effectiveness of computational analyses of texts, the methods for collecting, curating, and processing the dataset must be transparent, iterative, and thoroughly documented. However, methodologies and best practices regarding topic modeling and social network analysis of literary, and specifically figurative language data, require further refinement. Since establishing best practices, to a large extent, determines the degree to which we can depend on a project's claims of discovery and new knowledge production, this chapter

[63] According to John Sinclair's important work on the topic, "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." For more about creating digital corpra, see Sinclair, J. 2005. "Corpus and Text - Basic Principles" in *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2012-09-01].

contends with the practical and procedural aspects of collecting, curating, and preprocessing texts using figurative language for the purpose of topic modeling. Moreover, I suggest an iterative process of project development that attends to the fiscal and temporal demands of digital humanities work by producing short term models with lightly encoded texts that can evolve into more richly encoded datasets for longer-term project goals.

The story of the ekphrastic tradition, and women's relationship to that tradition, is in many ways the story of data collection and curation, as reflected by Mitchell's statement at the end of "Ekphrasis and the Other:"

> My examples are also canonical in their staging of ekphrasis as a suturing of dominant gender stereotypes into the semiotic structure of the imagetext, the image identified as feminine, the speaking/seeing subject of the text identified as masculine. All this would look quite different, of course, if my emphasis had been on ekphrastic poetry by women. But the difference, I would want to insist, would not be simply readable as a function of the author's gender. The voice and "gaze" of the male, as Williams's "Portrait of a Lady" should make clear, is riddled with its own countervoices and resistances, and no one is going to blame the Grecian urn for the banalities Keats forces her to utter….
>
> I have not mentioned the verbal representation of other kinds of visual representation such as photography, maps, diagrams, movies, theatrical spectacles, nor reflected on the possible connotations of different pictorial styles such as realism, allegory, history painting, still-life, portraiture, and landscape, each of which carries its own peculiar sort of textuality into the heart of the visual image. This treatment of ekphrasis, then, like the typical ekphrastic poem, will have to be understood as a fragment or miniature. (181)

What Mitchell points to at the end of his seminal essay is a human dilemma. It would be impossible to mention all of the kinds of ekphrasis, all of the "many figures of difference," that fuel his model of ekphrasis as a semiotic struggle staged within a tripartite network of speakers, artworks, and readers. Mitchell argues that while there

might be other "countervoices" and "resistances," as he calls them, within the ekphrastic tradition, the limitations of human reading and print publication force him to set them aside for later studies. Distant reading practices, however, offer promising alternatives to Mitchell's limited data collection—all poems by dead, Romantic or modernist white men.

Thus, the second part of this dissertation considers Mitchell's dilemma and responds to the human limitations of "Ekphrasis and the Other" by suggesting a distant reading methodology that, when combined with close, interpretive readings, offers a broader, more complex, and more inclusive alternative to the tradition and critical understanding of ekphrasis. By explaining the method and rationale for collecting thousands of poems, hundreds of which are ekphrastic, lightly curating them with descriptive metadata, and ensuring the best possible methods for preprocessing the corpora to answer questions relevant to our critical understanding of ekphrasis, this chapter sets the stage for a visualization and exploration of the dialectical relationships between ekphrastic poems and other ekphrastic poems, as well as ekphrastic poems and non-ekphrastic ones.

At the same time responding to concerns of the digital humanist, this chapter also demonstrates best practices for iterative project development that responds to the types of questions relevant to humanities scholars redressing issues of canon-formation and ekphrastic tradition. For example, through experimentation during the preprocessing of the data, I identify the stoplist[64] best-suited for exploring ekphrasis, and poetry more generally. For digital humanists refining their methodologies and for the literary scholar

[64] A stoplist is a file with high-frequency words in a language (English in this case) that are removed from a corpus of text before some form of textual analysis.

interested in uncovering latent patterns in ekphrastic poetry, the chapter that follows uses the refinement of digital practices to produce compelling results about the language of "stillness" and "looking" in ekphrastic verse.

Project Overview[65]

Ekphrasis offers a wealth of opportunities to ask familiar humanities questions about canon-formation, literary tradition, and genre definition, and at the same time affords avenues for the advancement or refinement of methods and tools in the field of digital humanities. Effective digital project design marries humanities questions with digital tools, algorithms, or other technologically-enabled processes to produce new knowledge, reveal latent patterns of language, or discover better questions.[66] The project described in this chapter and the following one strives to meet this goal by leveraging the inherent computational power of an algorithm called latent Dirichlet allocation (LDA)[67] to identify trends in ekphrastic texts for the purpose of discovering new ways of understanding the relationship between them. Such a comparison, a reading of relationships between texts in a corpus of hundreds of ekphrastic poems, would help to

[65] The data preparation, scripts for removing duplicates, scripts for extracting text and metadata, clean-up of text, preparation of texts to be imported into MALLET, configuration of the EC2 instance for MALLET experiments, and formatting of data exported from the MALLET model represent contributions from Travis Robert Brown. The generous contribution of his time and expertise has made this a much better project and chapter. Any error or misrepresentation of data, however, is solely my responsibility.
[66] For more digital humanities project design, see Daniel Pitti's "Designing Sustainable Projects and Publications" in Schreibman, Susan, Ray Siemens, and John Unsworth. Companion to Digital Humanities (Blackwell Companions to Literature and Culture). Hardcover. Oxford: Blackwell Publishing Professional, 2004 or Jeremy Bogg's blog posts on project development at "Digital Humanities Design and Development Process · ClioWeb." http://clioweb.org/2008/04/06/digital-humanities-design-and-development-process/ Web. 17 Sept. 2012.
[67] Kao, Anne, and Steve R. Poteet. Natural Language Processing and Text Mining. Springer, 2006. Print.

overcome the human shortcomings that Mitchell describes. LDA models[68] of thousands of poems refocus the question of ekphrastic tradition and tropes on the relationships between discourses.[69]

Reading at a distance affords scholars interested in ekphrasis a methodological alternative to semiotics or metaphorical comparison by detecting word frequencies, linguistic patterns, repeated phrases, or by detecting and predicting patterns across hundreds or thousands or even millions of examples. Rather than being limited by the human capacity to read a few texts at a time, distant reading practices facilitate the detection of subtle language trends across thousands of texts in minutes to hours. Strictly speaking, though, LDA is not a method of reading. LDA is a form of computer learning, an algorithm that through repeated iterations refines existing predictions about data in order to fine tune its accuracy. Most other distant reading tools, such as Many Eyes[70], Wordle,[71] or even many of the tools in TaPOR[72] depend on detecting frequencies of word use and patterns of repetition or analyze the linguistic patterns in text, much like the Stanford Natural Language Processing Group's CoreNLP.[73] Docuscope, another text analysis tool in the digital humanities, depends upon extensive lexica to categorize text

[68] Through a process that will be described later in the chapter, LDA produces a list of likely topics based on word distributions in a corpus of texts. More broadly, models are representations of a large concept, idea, or machine. In this case, a topic model represents the likely categories of language in a corpus of texts.
[69] The word *discourses* as it is used here is, perhaps, best defined by Melanie Kill in the glossary of Bawarshi and Reiff, 211. "Language in use and understood as participating in social systems so having determining effects in social life." This definition is particularly fitting because it is suited both to the purpose of discourse within a literary context of the poem, as well as within the social context found at the end of the chapter when the network diagrams place poetic language in groups.
[70] "Many Eyes." <http://www-958.ibm.com/software/data/cognos/manyeyes/> Web. 17 Sept. 2012.
[71] "Wordle." http://www.wordle.net/ Web. 17 Sept. 2012.
[72] "TaPOR: Text Analysis Portal for Research." http://portal.tapor.ca/portal/portal Web. 17 Sept. 2012.
[73] "CoreNLP." http://nlp.stanford.edu/software/corenlp.shtml Web. 17 Sept. 2012.

data.[74] However, LDA uses probability to refine its own methods of organizing and sorting data.

LDA, a form of probabilistic topic modeling,[75] therefore, presents opportunities previously unavailable for studying latent structures in poetic texts. Free verse and lyric poems are frequently at odds with the strict structure needed for semantic data mining tools; however, probabilistic topic modeling does not depend on correct semantic arrangements to work. Through Gibbs sampling, probabilistic topic modeling responds to the semantic ambiguity typical of figurative language by disambiguating words through samplings of other less polysemous words from the same document. Therefore, the same words employed in different contexts are parsed differently.[76] Sorting texts by the probability with which they include words that co-occur with similar words in similar texts renders "topics," which are groups of texts that the algorithm predicts share a proportion of common language.[77]

Considering the strengths of probabilistic topic modeling and the possible benefits of using latent patterns of language co-occurrences to ask questions about the canon,

---

[74] "Docuscope." http://www.cmu.edu/hss/english/research/docuscope.html

[75] A form of topic modeling that relies on advanced statistical models for predicting probability.

[76] For example, a word like "spot" could be used as a noun: "I saved you a spot." It can also act as a verb "Did you spot him?" Alternatively, it could also be a proper name: "Come, Spot!" LDA would parse the words differently based on context. The first example might appear with other words that have a likelihood of indicating location. The second might appear in a distribution of words indicating sight. The third might appear with a list of proper names. LDA does not determine the definition or meaning of the word. Instead it uses a form of probability to predict which other documents and the word "spot" is likely to appear.

[77] Chapter 5 will describe the LDA algorithm in much more detail. Matthew Jocker's soon-to-be published book, *Macroanalysis: Digital Methods and Literary History* (UIUC Press, 2013) promises to shed some light on methodologies useful for humanists interested in studies involving Latent Dirichlet Allocation (LDA). In the meantime, his blog post "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors" (http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/) offers a humorous, narrative introduction to the assumptions inherent in LDA models. Similarly, Scott Weingart's blog post "Topic Modeling for Humanists: A Guided Tour" presents an approachable introduction to LDA (http://www.scottbot.net/HIAL/?p=19113).

tropes, and tradition of ekphrasis at scale, I developed a digital project that collected over 4,500 plain text poems and created a modest metadata scheme to begin describing and curating the data to help interpret LDA models. Furthermore, I hoped to render the results of topic modeling with network graphs that facilitate interpretive and exploratory navigations through the corpus.[78] Using a tool called MALLET to run the LDA algorithm, I generated lists of topics from my private corpus of 19th through 21st century poetry, including non-ekphrastic as well as ekphrastic poems.[79] Furthermore, I reconsidered assumptions about ekphrastic poetry in light of the topic distributions produced by the model. For example, one might expect that a topic including words about stillness and muteness might be the most common topic in ekphrastic poetry because theories of the genre take as a given ekphrasis' reliance on the binary tension between word and image, time and space. Similarly one might expect another topic to form around the language of rivalry, which represents what recent scholarship calls an over-determined feature of the genre. Therefore, I used the following three questions to guide the selection and preparation of the dataset used in the topic modeling experiments.

1.) Could a computer distinguish differences between poems by men and by women? In "Ekphrasis and the Other," W.J.T. Mitchell argues that were we to read ekphrastic poems by women as opposed to ekphrastic poetry by men, we might find a very different relationship between the active, speaking poetic voice

[78] This project was funded in large-part through a fellowship from the Maryland Institute for Technology in the Humanities (MITH) http://www.mith.umd.edu/. The technical, data curation, and preprocessing described herein are informed by the generous collegiality, time, and thoughtful conversations with MITH's staff, especially Travis Brown, Trevor Muños, and Jennifer Guiliano.
[79] MALLET's: MAchine Learning for LanguagE Toolkit." http://mallet.cs.umass.edu/. Cameron Blevins makes the case with regard to modeling Martha Ballard's diary, which is later revised by Clay Templeton in a MITH blog entry suggesting that MALLET's ease of use makes it the best "out of the box" program for humanists. See http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/.

and the passive, silent work of art—a dynamic that informs our primary understanding of how ekphrastic poetry operates. Were this true and were the difference to occur within recurring topics and language use, a computer might learn to recognize latent textual patterns more likely to occur in poetry by men or by women.

2.) What role does the language of stillness play in the latent patterns of ekphrasis? Would topic modeling of ekphrastic texts pick out "stillness" as one of the most common topics in the genre? Much of the definition of ekphrasis revolves around the language of stillness: poetic texts, it has been argued, contemplate the stillness and muteness of the image with which it is engaged. Stillness, metaphorically linked to muteness, breathlessness, and death, provides one of the most powerful rationales for an understanding how words and images relate to one another within the *ut pictura poesis* tradition—usually seen as an hostile encounter between rival forms of representation. The argument to this point has been made largely on critical interpretations enacted through close readings of a limited number of texts. Would a computer designed to recognize co-occurrences of words and assign those words to a "topic" based on the probability they would occur together also reveal a similar affiliation between stillness and death, muteness, even femininity?

3.) Could a computer detect vocabularies or combinations of words and images that distinguish poems as ekphrastic or non-ekphrastic? Mitchell explains that "no special textual features can be assigned to ekphrasis, any more than we can, in grammatical or stylistic terms, distinguish descriptions of paintings statues, or

> other visual representations from descriptions of any other kind of object" (159). We base this assumption on human, interpretive, close readings of poems; however, there is the potential that a computer might recognize subtle differences as semantically significant when considering hundreds of poems at a time.

In general, these are small questions constructed in such a way that there is a reasonable likelihood that we may get useful results because they draw on the algorithmic strengths of probabilistic topic modeling.[80]

Furthermore, the current chapter demonstrates a project design process that mirrors iterative design principles from the computer sciences in order to produce a project that is actionable in the short term, sustainable in the future, and responsive to the evolving needs of the project in order to improve, expand, and enrich the project's findings over the long term. The next three sections of this chapter present the methods, rationales, and future goals of three important aspects of the project that determine the efficacy and reliability of chapter 5's LDA tests. Through each of the following sections, I strive toward transparency about the dataset, its collection, its curation, and the preprocessing techniques that prepare it for textual analysis. Like most digital projects, the dataset, metadata, and preprocessing techniques will continue to evolve and improve. Each of the following sections will describe how changes, errors, and difficult choices about selection and processing participate in ongoing improvements to the project that

[80] The choice of the word "results" instead of "answers" is purposeful because none of these would be answers. Instead the result of each study is designed to turn critics back to the texts with new questions.

require documentation and version control.[81] The advantage to LDA models is that they do not require extensive meta-tagging to produce salient results.

The Data

Text mining and computational tools designed to analyze latent patterns in natural language data present researchers interested in contemporary literature with daunting challenges. The question of how to acquire a large set of already electronically available poems from the twentieth-century on confronts the challenges of copyright protection and availability. Currently, there are no existing public datasets of contemporary poetry as there are for literary works published prior to 1922. The lack of public, digital collections of contemporary literature available for humanistic digital and computational analysis prompts Mark Sample to ask: "how have scholars of contemporary American literature been left behind by the rise of digital tools and the methodologies afforded by those tools that have otherwise been a boon to literary scholars working on earlier eras of American literatures?" (*Debates* 188) Sample points to the constantly-extending length of copyright restriction that prevents researchers from accessing, using, and publishing from digitized texts. He continues:

> Although it's risky to generalize about the digital humanities, it is safe to say that the work of the digital humanities is ultimately premised upon a simple, practical fact: it requires a digital object, either a born-digital object or an analog object that has been somehow scanned, photographed, mapped, or modeled in a digital environment. In the context of literary studies, this usually means a large corpora of digitized texts, such as the complete works of Shakespeare, the multiple versions of Whitman's *Leaves of Grass*, or every single book published in England during the nineteenth century. (188)

[81] Version control is to information science and data curation as variorum is to texts. In this project the repository where each version of the data is saved is called GitHub.

While he specifically highlights the disadvantage that copyright restrictions cause for the study of latent and predictive patterns in Don DeLillo's work, Sample argues that DeLillo is really a metonym for contemporary authors whose work will remain under lock and key for years to come, untouchable in the form of large, public archives, which have fuelled (and continue to) the explosion of digital humanities work in literary period studies prior to 1900. Admittedly, collections such as HATHI Trust *do* include collections full-text, searchable copies of some contemporary works; however, the results return only page numbers corresponding with searched text. Access to HATHI Trust full-text files is only available through special research arrangements.

How, then, can we leverage the power of tools for text mining, pattern recognition, and corpora discovery for the purpose of studying texts written after 1922? Could private, unpublished collections of modern poems yield a collection that would be sizeable enough to study using LDA algorithms and still produce salient results? Drawing from existing Web sites that publish modern verse may push the boundaries of what is acceptable in terms of copyright restrictions; however, in the short term the ability to use the data, even if none of it could be published, offered a promising mid-way solution. In the short term, I created a private collection of poems drawn from online, public resources for private research use, allowable under copyright and fair use law. Doing so allows me to test algorithms, but this solution means that the only data I can

make available to readers is the data produced by the model that does not present the possibility of reassembling whole texts.[82]

Therefore, the data set used to perform the following tests and studies, like many digital humanities projects, reflects not just the research agenda of the project as a whole, but the particular practical and editorial constraints of what is available. In order to perform a computational analysis of thousands of poems, there needs to be an electronic resource with digital files/copies of thousands of poems. When working with large data or small data using computational text analytics, one clear challenge from the outset is finding the right dataset to work with.

Funding for the project extended for only four months; therefore, scanning poems from print sources with optical character recognition (OCR) software risked spending too much time on data collection to the exclusion of the project's other goals. The next best option was to collect electronic copies of poems from online, public content providers, such as *The Academy of American Poets'* Web site (poets.org) and the *Poetry Daily* Web site (poems.com) to create a private digital repository of mostly twentieth-century poems. *Poetry Daily* (poems.com) is an online anthology of contemporary poetry. Designed to publicize the most recent work produced by contemporary poets, *Poetry Daily* reproduces one poem (and sometimes two) each day. The site displays each poem on its site for exactly one year. After one year, the online version of the poem is retired. Given the one-year agreement with literary magazines and small presses, the poems in *Poetry*

[82] Topic distributions of key words produced by LDA models in this study consist of single words, and therefore cannot be used to reproduce the original text. In the short term, the only data that I can publish includes the titles of poems and the statistical data generated by the model; however, in future iterations of the project, I hope to collaborate with content providers, exchanging their data for my results to improve their site's overall navigability and my project's access to more data to build and refine future LDA models.

Daily's online anthology are often published within the past five years. The most recently published poems in the digital collection used for this study come from *Poetry Daily.*

The overall corpus, comprised of 4,771 documents, was assembled from five sources by using a macro[83] to produce digital copies of poems as individual documents, including the title of the poem, the name of the poet, the text of the poem, and any available publication information, including the name of the book or journal where the poem was originally published, the date, and the name of the publisher.[84] The largest content provider, *The Academy of American Poets* Web site (poets.org) generated 4,266 total poems for the collection, and *Poetry Daily* (poems.com) added 373 poems to that. Another portion of the poems in the corpus includes specifically ekphrastic poems from popular print anthologies or bibliographies. These poems were often keyed by hand. In particular, 34 poems came from John Hollander's anthology *The Gazer's Spirit* ; 79 poems were discovered from Robert Denhem's *Poets on Paintings: A Bibliography*. Additionally, I added to the corpus ekphrastic poems by women whose work was not included in the aforementioned sources, including poems by Jorie Graham, Carol Snow, Barbara Guest, and Cole Swenson, thus accounting for the remaining 19 items. Table 4 breaks each source and number of poems down, including the proportion of poems from each source.

[83] A simple program written to accomplish repetitive tasks. The scraping macro used to collect electronic copies of poems, written specifically for a Mozilla Firefox plug-in called iMacro, opened each Web page, selected the title, author, poem text, and available publication information and dumped each one into its own plain text file.
[84] The data collected from *Poetry Daily* represents those posted on the site between January 15, 2011 and extending through January 14, 2012.

Table 4: Total number of poems from each content source and percentage of the corpus comprised of each source

Source	Total # of poems	% of corpus
Poets.org	4266	89.4%
Poems.com	373	7.8%
The Gazer's Spirit	34	.7%
Poets on Paintings	79	1.7%
Graham, Snow, Guest, & Swenson	19	.4%

In an effort to keep the poem data as clean as possible and free from error, a second script removed all of the metadata from each text file, leaving only the title of the poem and the body of the poem in plain text. The digitized text features of the poem were made UTF-8 compliant.[85] Next, each document was assigned a unique identifying number. Poems from poets.org received the prefix po- followed by a 6 digit number. Files extracted from poems.com received a pd- prefix followed by a 5 digit number; likewise, files from John Hollander's *The Gazer's Spirit* begin with the prefix gs-. Documents with poems from *Poets on Paintings* and the small group of women poets received the prefixes respectively: fc- and sg- followed by a 6 digit number. Randomly generated, the numbers became document identifiers as well as the name for each poem file. The entire collection of poems resides in a directory specifically for the corpus's text files, divided based on the source of its collection. Changes to the collection during the data standardization processes were tracked in a private GitHub[86] repository.

We seem almost inherently to know the value of "big data:" scale changes the name of the game. Still, what about the smaller universes of projects with minimal

[85] UTF-8 stands for USC Transformation Format – 8 bit. This is a universal format for encoding digital text.
[86] GitHub is a repository for the data equivalent of text variorums.

budgets, fewer collaborators, and limited scopes, which also have large ambitions about what can be done using the digital resources we have on hand? Without detracting from the import of big data projects, smaller projects offer the field rich opportunities for exploratory studies using advances in natural language processing tools, and the outcomes of such projects can be relevant and useful both in and of themselves as well as beneficial to large-scale projects by providing possible methods for tasks such as fine-tuning initial results. Small data sets such as this one prompt digital humanists to ask questions like: how do we recognize useful results? How do we know if our algorithms are working the way they are intended? Trying to answer questions about metadata curation, interoperability, and detail with big data can be expensive and time consuming, but small and mid-sized data sets can be more deeply and inexpensively encoded. Herein lies the necessity for discussing my methods, as in this chapter, alongside results—the topic of chapter 5. Methodological documentation is as important to the digital humanities as the refinement of theoretical concepts has been to the study of literature.

Importantly, small projects (and even mid-sized projects with mid-sized datasets) offer the promise of richly encoded data that can be tested, reorganized, and applied flexibly to a variety of contexts without potentially becoming the entirety of a project director's career. The space between close, highly-supervised readings and distant, unsupervised analysis remains wide open as a field of study, and yet its potential value as a manageable, not wholly consuming, and reproducible option make it worth seriously considering. Small to mid-sized data collections are often flexible enough that an iterative project design process allows frequent improvements and refinements to the data collection that can be seamlessly folded into the data versioning and project development

process. For example, in a small corpus, additions to the corpus can influence the results of the entire study more easily. Furthermore, data corrections are much easier to accomplish and can happen much more quickly in response to confusing results or test errors. By maintaining clear records of the evolving state of the data through a version control system such as GitHub, small projects can more flexibly respond to improvements, adjustments, and refinements of the dataset that help better address the humanities questions the project is attempting to ask. Results of data mining experiments with small datasets can also be more easily interpreted in light of subject area expertise. Because the project is small, future studies will likely focus on how adding and removing items from the dataset influences LDA results. New iterations of the project can develop quickly as the dataset grows and that the number of ekphrastic poems in the collection increase, thereby improving the reliability and scope of the project as a whole.

The promise of iterative design in small digital humanities projects is that we can begin to build, test, and produce initial results while at the same time refining, improving, and expanding the data, metadata, and preprocessing techniques. In terms of the data collection process, the need to produce a substantive enough dataset also required some compromise in terms of the kinds of data captured. For example, harvesting digitized poems from the Web was not a perfect solution. In an effort to capture the most reliable online resources and by including all the poems from poets.org, the data set includes a undefined number of poems published before 1900, and our ability to define precisely how much of the dataset consists of pre-1900 poems depends on a much lengthier process of metadata formatting and curation. Further drawbacks to using these two electronic collections include the lack of transparency regarding editorial selection. Both

collections are largely assembled based according to editorial preferences that are not clearly stated on the Web site.[87] Rather than representing my own choice of poems, I must rely on the selection of texts, editions, and textual variances that fit the site's editorial preferences. Furthermore, poems from *Poetry Daily* will vanish after 365 days, making corrections or references back to the original virtually impossible.

Though the presence of pre-1900 texts may detract from the reasonable claims to be made with regard to periodization, the implementation of short-cycle, iterative design in which version controls track changes to the dataset, suggests that small bursts of human and programming interventions to improve the dataset have the potential to make substantive improvements in the overall project. Moreover, the existing dataset allows us to begin modeling a corpus of poetic texts right away and ask questions about the model's outcomes that respond to issues of poetic tradition, tropes, and genre definition.

Metadata

One of the advantages to using LDA is that it does not depend upon a richly-encoded set of metadata—data describing data—to produce salient results. On the other hand, LDA, which is an unsupervised form of data mining, generates descriptive metadata that can be used for navigation and exploration. However, coupling even small amounts of metadata with LDA studies creates richer conditions for using LDA as an exploratory, as well as a descriptive, tool.

Metadata quality, standards, and curation are concerns close to the hearts of librarians, but metadata is evolving as an important consideration for literary scholars

[87] Unfortunately, a written request for Poets.org's editorial priorities and policies has not yet been responded to by the site's publisher.

because it helps us to organize and develop our hermeneutic approach toward computational analysis. In "Metadata for Corpus Work," Lou Burnard argues that particularly in the case of working with linguistic analyses of digital corpora, metadata plays a central role in understanding and interpreting test results.[88] He writes:

> Nevertheless, it is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible. Why? Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language. A typical corpus analysis will therefore gather together many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata restores and specifies that context, thus enabling us to relate the specimen to its original habitat. Furthermore, since language corpora are constructed from pre-existing pieces of language, questions of accuracy and authenticity are all but inevitable when using them: without metadata, the investigator has no way of answering such questions. Without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity.

Granted, LDA, which is a probabilistic modeling algorithm rather than a linguistic one, depends less on the pre-existence of rich metadata because the model produces its own metadata that, as we will see in chapter 5, helps "restore and specify" context. However, the point should not be overlooked. Lightly curated metadata accompanying the textual data in combination with the metadata produced with the model allows for better visualization of test results and increases the number of interpretive options.

Consistent and accurate metadata benefits interpretations of the model in three ways. First metadata helps expose our human assumptions and biases about the dataset. By producing only two categories of metadata—the gender of the poet and the genre classification of the poem—we can explore ways in which traditional definitions of

[88] More on metadata standards and creation can be found in Burnard, Lou. 2005. "" in Developing Linguistic Corpora: a Guide to Good Practice , ed. M. Wynne. Oxford: Oxbow Books: 30-46. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2012-09-05].

ekphrasis compare to the model's predicted topic distribution. Secondly, accurate metadata allows the researcher or research team to supervise the model creation process and provide nuanced oversight over the algorithm. For example, in some topic models where trends over time are particularly significant, metadata creates the conditions under which textual features can be mapped over time. In point of fact, Rob Nelson's study of the *Richmond Daily Dispatch* tracks strains of nationalistic language (such as elegiac or celebratory) over the duration of the Civil War. Nelson's chronological graphs demonstrate correlations between elegiac and nationalistic language with casualty and enlistment rates. As a result, Nelson can make arguments about the effectiveness of particular forms of nationalistic rhetoric over time and in response to corresponding calls for enlistment and reports of war fatalities—all of which becomes possible because of a richly encoded data set. Thirdly, metadata improves the type of data and results that can be shared about a copyright protected or private data collection. In cases such as this project, when the full text is unavailable for use by readers, metadata created by the researcher can be made available in its place.[89] Since the majority of documents included in my corpus are still under copyright protection, I cannot make the original, plain text data visible to readers and future researchers; however, I can make metadata I generate available. Finally, metadata turns individual files into collections, articulating a collective purpose through the standardization of description that, when well formed, also clarifies its intended use, by tailoring metadata to the questions being asked and the desired outcome of the results.

[89] Substituting metadata for content is a solution commonly offered by content repositories and libraries with regard to electronic distribution of copyright-protected materials.

Even for small projects, developing standardized metadata is a time-intensive process. The decision to create metadata for this digital collection, then, represents a deliberate choice to balance what might be done with what can reasonably be accomplished in this lifecycle of the project. In his introduction to a 2009 issue of *Digital Humanities Quarterly*, Matthew Kirschenbaum poses the following:

> How do we *know* when we're done? What does it mean to "finish" a piece of digital work? As Bill Kretzschmar points out in his essay, the verb "to finish" can mean to *complete* or something more like to polish or perfect. What is the measure of "completeness" in a medium where the prevailing wisdom is to celebrate the incomplete, the open-ended, and the extensible? (2)

Digital projects—due in part to the plasticity of their digital environment and in part to the innovation-centric funding mechanism that rewards new beginnings over completeness—rarely end, unless they are forgotten or set aside and never in that instance considered "done." Those are more likely to be "archived" as half-made proofs-of-concept fallen short of expectation. At first, iterative project design principles may appear to perpetuate the open-endedness Kirschenbaum describes; however, projects that are designed to respond to exigent humanities questions with small datasets and that begin to address those questions early in the project's development are more inclined to sustain interest and support. Iterative projects may never be done, but they can be over when they evolve into something else.

Embracing iterative design principles, this project begins by adding a small amount of metadata that can be added to incrementally, fuelling short-term discoveries in a timely way while at the same time staging improvements that increase or extend the project's long-term possibilities. Metadata created and used in this study (and available

in Appendix A), targets specific research questions about latent trends in ekphrastic discourse that break down along gender lines, latent patterns in language surrounding stillness, space, and femininity, and latent features of language that distinguish ekphrastic from non-ekphrastic poems.

Standards for capturing, producing, and standardizing metadata continue to be an ongoing conversation in digital humanities communities that use probabalistic topic modeling. Despite a disciplinary tectonic shift toward establishing metadata standards for interoperability[90], no particular best practices have been articulated to this point for creating and building small, private corpora with metadata intended for use in topic modeling. Consequently, the metadata creation and curation process for this project combines broadly-conceived best practices for metadata standards with the practicality of what works.

When the poetry data was collected, descriptive information about each poem was stripped (copied) out of individual text files with a script and placed into corresponding fields in a spreadsheet. For example, all of the author, title, and publication data (such as book title, publisher, and year) were extracted from the text file and placed into separate fields of the metadata spreadsheet. Repeating this process consistently across every file during the collection process, the descriptive information, called metadata, populated a spreadsheet with the following fields: document identifying number (doc id), title (title), author (author), publication information (pub info), notes (notes). Since the Web pages from which the poems were extracted were not always consistent about formatting or

[90] The ability of data to work between various kinds of technology. For example VHS and BetaMax were not interoperable, leading to the demise of the Betamax format; however, html pages are more interoperable, as they can be read on a variety of browsers.

including publication data, anything that might be used to describe a poem's publication history was combined and entered into a single field.[91]

Adding to and modifying the metadata in spreadsheet form[92] improved the process of entering and refining metadata by simplifying and regularizing it. Features such as auto fill improved my ability to group segments of data and enter the metadata more efficiently. With the use of simple processing scripts, the comma-delimited spreadsheet could be used in the future to generate a TEI header.[93] Furthermore, using Excel helped keep field and category names consistent. I particularly liked the flexibility of modifying and editing metadata in a spreadsheet and then exporting the data into whichever standard—RDF or TEI header—was needed later.

In a concentrated effort to build consistent metadata, I established criteria to help make consistent choices. Specifically in the case of gender assignment, I assigned each author to the categories of male, female, or unknown, based on the best available data (such as author or biographical statement) from online collections, print anthologies, or other scholarly sources, such as journal articles or biographies. If it was difficult to ascertain gender from the given information, the gender category was tagged as unknown. In the event that a particular author's gender self-definition changed between the publication of one poem and the publication of another poem, the gender definition at the date of poem's publication was used, or, as in the case of opposite-gender

[91] This process, as it turns out, was not without complication, as some publication fields included commas, semicolons, and other markers that later introduced errors into the dataset; however, the errors introduced do not immediately impact the effectiveness of the LDA model. Instead, correcting this data field became one of the decisions against perfection and in favor of short-goal completion. Correcting publication data, then, will become part of the next iteration of the project.
[92] Microsoft Excel, 2010.
[93] TEI is one recognized schema for encoding documents with descriptive information. TEI – Text Encoding Initiative, http://www.tei-c.org/index.xml Web. 10 Sept. 2012.

pseudonyms, the gender-definition of the author's writing persona was used (as in the case of George Eliot).[94] Among the 1,868 poets whose work is represented within the corpus, 681 are women and 1,021 are men. The remaining 165 authors' names are tagged "unknown" including authors who purposefully obscure their gender identity, published anonymously, or were part of a 10% sample designated for a classification analysis. Of the poems collected from *The American Academy of Poets*, 1,360 are by women and 2,570 are by men, and 334 poems were marked "unknown." Among the poems collected from *Poetry Daily* approximately 142 of those poems are by women, while 192 are by men. The remaining 37 poets are tagged "unknown."

In an effort to ensure consistent genre category assignments, I defined the criteria a poem must meet before being tagged as ekphrastic at the outset of the project.[95] Having to assign clear labels to poems often tested the resolve of the sometimes arbitrary-feeling decisions between what makes one document ekphrastic and the other not. However, each decision represented a conscious effort to remain consistent. If a poem included the name of a work of art, obvious description, paratextual information on the document's Web page, or was taken from an anthology of ekphrastic poetry, I labeled it ekphrastic. Poems that mention artists' names, but that did not mention a specific work of art were labeled "unknown" unless or until a secondary source, such as a syllabus, collection, or

[94] There are a few exceptions to this rule, however. In earlier stages of the project's development, I hoped to use gender and genre categories to train a classification algorithm to test if it could distinguish differences between ekphrastic poems by men versus those by women. In order to create that study at least 10% of the data needed to be classified as "unknown" in order to run those tests. As a result, some records were intentionally left tagged "unknown."

[95] In order to refine the criteria for ekphrastic, non-ekphrastic, and unknown category assignments, I began with a random sample of poems using a random number generator. Reading those poems, I considered what was a reasonable degree of research that could be performed in order to finish the metadata creation in a timely way. The genre definition guidelines were derived from recording criteria that would correctly describe most of the 20 poems in my initial sample.

scholarly article cited a specific work of art to which it referred. Notional ekphrasis, those poems that create imagined works of visual art, such as Keats' Grecian urn, were labeled ekphrastic so long as at least one other source also recognized the poem as ekphrastic—for example, it was included in the "Poetry about Art" section of *The American Academy of Poets* Web site or found in an anthology of ekphrastic poetry. Poems that included one or more easily-recognized ekphrastic trope, such as possible envoicing of an art object or an epigraph including a dedication to a painter, or poems that include extended descriptions of domestic objects that might be considered ekphrastic (such as a map or a bowl) were also categorized as "unknown." Finally, poems that did not meet either the criteria for "ekphrastic" or "unknown" were classified as "non-ekphrastic." All in all, 276 poems in the entire corpus were tagged ekphrastic. Future iterations of this project will continue to increase the number of ekphrastic poems in the collection.[96] In the following chapter, we see two ways this metadata fuels how we select, model and interpret the data through the LDA process; however, it should be noted that the usefulness of the digital collection and the model stems from its inclusion of both ekphrastic and non-ekphrastic contributions.

Metadata concerning the existence of duplicate poems in the database improves the reliability and efficacy of the model's results. During the process of creating the dataset, duplicate files resulted from one of two possible avenues.

[96] Future work would also include a classification analysis, which unfortunately was not something that could be accomplished during the grant-funded portion of the project. Classification analysis includes training an algorithm on a set of exemplary data (a combination of ekphrastic and non ekphrastic) and then using the computer's decision trees to predict the classification of an "unknown" set of data. By parsing through classification decision trees, one might identify the algorithm's learning process and begin to ask questions about how human assumptions about methods of classifying ekphrasis may be understood by contrast.

1. A file may appear in the original poets.org or poems.com database twice. These were the easiest files to catch, because they are most often identical in title and author formatting.
2. A file may appear in multiple sources under slight changes in title and author name. In other words, if William Carlos Williams' "Musée des Beaux Arts" appeared in *The Gazer's Spirit* edited by John Hollander, which it does, and was keyed in without the accent mark, but it also existed in the poets.org database, but generated an error during the conversion to text where the accent mark had been. It is possible that the script designed to tag duplicates read these two files differently and therefore did not mark them as duplicates.

All in all, precautions to capture duplicate files generated an additional metadata category, marking every file in the database after the first instance as a duplicate. When the text data is merged with the metadata during the preprocessing stages of the LDA tests described in the next chapter, files marked as duplicates are left out of the dataset used for LDA analysis. This results in a decrease of files from 4774 to 4500. Despite our best efforts, some duplicates that were not detected by the duplicate script remain in the data set when it is imported into MALLET; however, these files do not seem to pull the topic distribution in ways that decrease the accuracy of the model. In fact, in some instances, the ability of the model to assign nearly identical topic proportions to duplicate files serves as further indication that the model is working—while this may seem foreign now, this concept will become clearer in the next chapter. Future studies using this dataset would benefit from careful and perhaps manual data cleaning that fixes special characters and repairs their UTF-8 assignment.

Iterative design, version control, and purposeful imperfection

Iterative design processes require establishing clear methods for version control for primary data, for metadata, and for LDA modeling data. As digital projects are seldom solitary undertakings, version control repositories such as GitHub[97] allow multiple users to read, modify, and track versions of their data in a way that can be recovered if needed either because the existing dataset was corrupted or because there is a need to re-run an earlier test on a prior version of the dataset. The advantage to continuing to use GitHub to track versions of the data and metadata is that the project's data set and meta-data can continue to be improved, updated, and increased, while at the same time establishing a record of the project's legacy data. Through version control and careful record keeping, changes to the data can be introduced in the middle of the project to produce immediate results.

Why should such a thing matter to the literary scholar? If, as I have argued throughout this dissertation, our understanding of the ekphrastic tradition is a dialectical process of discovery, engaging with historical tropes, and readjusting, then the ability to compare future data models with those that are more limited or biased represents the enormous potential for future scholarship. Versioning datasets allows researchers to compare their evolution, leading to a better understanding of how the data, the humanities questions, and the rich points of inquiry have changed over time. Similarly, if metadata is a form of criticism, descriptive or predictive, it underscores the existing beliefs of the researcher who created it.

[97] GitHub is the repository in which the data versions are held <https://github.com/>.

As the earlier quotation from Kirschenbaum makes clear, data and metadata are elusive aspects of digital projects. None is ever big enough, clean enough, or well-structured enough to achieve precisely what it is that researchers would like to achieve. Just at the point where the "perfect" dataset seems within reach, new standards or technology are released, creating new needs or even opportunities to create "the perfect dataset." Most projects have lifespans determined by fellowships or grants or sabbaticals, and we can't afford to spend the entire project chasing a standard that simply doesn't exist. In fact, the DH mantra may well be "project or perish." Hard decisions about data formatting and metadata creation are often determined by two factors: intellectual value and time. First, data should be thoughtfully selected, described (tagged with metadata) and well-formatted enough in order to work and to reasonably make the argument that your results can be trusted. The best and highest-value data and metadata address the central questions of your project, and when they are not absolutely necessary, it is time to make the difficult decision to stop. To engender the values of iterative project design often means choosing between good-enough and great, and creating the data and metadata for this iteration of the project required that choice. Focusing on gender and genre for metadata categories, I decided to let go of other tags, like publication date or URLs pointing to online and freely available copies of the poems in the dataset; however, by choosing not to pursue perfection, the project continued on schedule.

Each missed opportunity, excluded poem, and unremarked upon metadata tag presents opportunities within the small project environment to make substantive improvements in short, serialized steps. Perfect data sets are a myth, one that often forms a barrier to scholars who wish to begin DH projects and feel surrounded by projects

"more perfect" than what feels achievable at first. Rather than struggling for the perfect data set, I want to suggest that we place a much stronger emphasis on the more intellectual and more necessary component of data curation—data versions. I would argue that we judge projects not by the "completeness" or "perfection" of the data, but how well its versioning has been documented, how thoroughly curatorial decisions such as what to tag, when, and why have been publicized and described, and how much the evolution of the data contributes to the development of other projects within the community. In much the same way that we know more about the value of an article by how often it has been cited, we should value a digital humanities project by how much can be learned by the projects that follow it.

The Story of Stopwords; or, Data Preparation

In the previous two chapters, close readings of ekphrastic poems by Elizabeth Alexander, Lisel Mueller, and Elizabeth Bishop take seriously the influence so-called "little" words like "to," "by," "on," and "above" have in the poem's semantic composition. Distinct for its highly concentrated language, poetry places an increased degree of significance on even the poet's "smallest" word choices. For example, Robert Frost's poem "Stopping by Woods" would be a different poem if it were titled "Stopping in Woods," and the difference would be more than a matter of the speaker's physical proximity to the woods—near rather than in them. The change would resonate aurally, as well. Consider, then, the difference that would be made by removing punctuation, line breaks, diacritical marks, capitalization, and high-frequency words. Auden's opening line "About suffering they were never wrong, / The Old Masters: how well they understood…" would be dramatically different if instead it read: "suffering never wrong

old masters understood." What if Keats' Ode opened: "unravish d bride quietness foster child silence slow time sylvan historian express flowery tale sweetly rhyme"?

The first steps in preparing documents for topic modeling require making such changes. Preprocessing strips documents of upper and lower case letters, removes line breaks and enjambments, deletes high-frequency words including articles, prepositions, pronouns, conjunctions, and common verbs—like "is," "are," and "were"—and turns documents into strings of sequential words that no longer bear the same syntactical meaning they once did. Given this, how can a methodology that requires radical decomposition of a poem's linguistic meaning offer valuable insights into exploring texts? Knowing, as we do, the significance of the line, "She looked on, and her looks went everywhere" in Robert Browning's "My Last Duchess," most literary scholars would be and should be nervous about the fact that preprocessing removes the whole line in its entirety from the text of the poem. Each word in the duke's pivotal line "justifying" his implied action is included in a list of words, called a stoplist, stricken from the text before it is imported into the LDA environment. In topic modeling, the words "she" "looked" "on" "and" "her" "looks" "went" "everywhere" exemplify frequently used lexia that hold little semantic weight. Because of their frequency, however, prepositions, articles, conjunctions, and other high-frequency words are removed from the corpus so that their sheer volume does not skew the results of the LDA model. While the understatedness of Browning's line actually underscores its significance within the poem, the same line in, say, the middle of a transcript from a congressional hearing would not operate the same way, and LDA algorithms were developed with non-fiction prose, not poetry, in mind. Raising these issues illustrate how different one methodology (topic

modeling) might be from another (like close reading), but more importantly require understanding how the deformative[98] aspects of preprocessing first determine the creation of LDA topics and the model's predictions about likely similarities between documents and then influence whether and how interpretations and discoveries can be made with the model's output. The results, surprising as they were, emphasize the promise and interdependence of close and distant reading practices as cooperative methodologies.

As I noted earlier, the ekphrastic networks created in chapter 5 use a computer program called MALLET to create LDA models of text corpora by predicting the likelihood that documents with similar patterns of language use are most likely also close thematically. What precisely "topics" are will be considered in closer detail in the following chapter; however, before the LDA algorithm in MALLET is possible, the texts in the corpora need to be converted into a format MALLET can understand. Over the course of the final pages in this chapter, I review the decisions made during the preprocessing stages of the topic modeling experiments that form the basis of networked readings of ekphrasis. As we will discover along the way, preprocessing—done in a purposeful and reflective way—leads to a discovery about the language of stillness in ekphrasis and the ability for machine learning techniques to detect stasis with the articles,

[98] I purposefully invoke the term *deformance* here to call to mind Jerome McGann and Lisa Samuel's use of the term, in "Deformance and Interpretation" in which they write: "A deformative procedure puts the reader in a highly idiosyncratic relation to the work. This consequence could scarcely be avoided, since deformance sends both reader and work through the textual looking glass. On that other side customary rules are not completely short-circuited, but they are held in abeyance, to be chosen among (there are many systems of rules), to be followed or not as one decides. Deformative moves re-investigate the terms in which critical commentary will be undertaken. Not the least significant consequence, as will be seen, is the dramatic exposure of subjectivity as a live and highly informative option of interpretive commentary, if not indeed one of its essential features, however neglected in neo-classical models of criticism that search imaginative works for their "objective" and general qualities."
http://www2.iath.virginia.edu/jjm2f/old/deform.html

conjunctions, prepositions, and high-frequency words such as "still," "stillness," and "say" removed from the poems' text.

I admit that I began as a non-believer—none of my training as a literary scholar or reader of poetry prepared me for the fact that the little words, upon which so many papers and articles in literary studies hinge, could be completely removed from a text and yet produce results I could trust. Removing high frequency words from the collection of poetry does affect the outcome of a computational analysis, but the results were not at all what I expected. Despite my vehemently held belief that high-frequency words carry more semantic weight in poetry than in prose, the stopword tests prove that we can produce useful and reliable results without them. Additionally, the results show that stillness and stasis in ekphrasis are more evident when the direct references to them are removed and their co-occuring metaphors and language divide into multiple, diverse discourses, ranging from peaceful to anxious. Although computational studies of literary texts and the close reading practices of literary scholars appear at first glance to be contradictory when it comes to how we read, the two fields in combination can work to the advantage of literary scholars by refocusing and occasionally distorting the lens of close reading to bring the latent patterns of texts into clearer focus.

Despite digital humanist's celebration of MALLET as a robust but approachable tool for topic modeling, few humanities projects consider how its preprocessing steps affect its output. While some digital humanities scholars, including Matthew Jockers and Ted Underwood, opt to write their own custom LDA programs using a programming environment like R, the choice to author one's own LDA modeling program is neither compelling nor practical. MALLET provides a robust, extensible, and perfectly viable

solution, and choosing to use tools that already exist and work well saves precious time during the lifecycle of a project. MALLET made the most sense for this project because it was relatively easy for me to learn and consequently would make my results easier to share. Furthermore, by choosing to use MALLET, my discoveries, trials, and results can be more readily applied by other scholars in the short term, thereby reducing the need to learn how to program rather than increasing it. For DH methods to become more commonplace, we must reduce the threshold of understanding and acceptance rather than unnecessarily increasing it.

In the field of computer science, where methods and algorithms of natural language processing (like LDA) were developed, high frequency words that hold little semantic weight create a high noise to signal ratio. In most applications of algorithms such as LDA, high-frequency words overwhelm the model, skewing it away from semantic clusters. The sheer repetition of articles, conjunctions, prepositions, common transitive verbs (was, is),or simply common verbs (look, say, see) overshadows less frequent but semantically weightier words. To correct for this imbalance, developers of natural language processing algorithms compile lists of high-frequency words that are routinely removed from the dataset. MALLET is no exception. The default settings for importing data into the program removes stopwords listed in Appendix B. In practice, stopword lists improved the results of LDA algorithms on large corpora of non-fiction texts such as grant proposals, *Science* magazine, and Congressional testimony.

To a computer scientist, this all seems quite obvious. If you are a literary scholar, particularly if your object of study is poetry—the previous paragraph is likely to be anxiety producing. The "little" words removed by the MALLET default list, we know,

are not filler words. They may not carry semantic weight, but they do carry syntactic significance, which is to say that they often determine or define the semantic weight of the words surrounding them. As poetry critics, readers, and writers know, articles, conjunctions, prepositions, and pronouns are all an important part of an art form valued for its economy of language. In fact, some of the most interesting articles in literary studies hinge on such things. Consider the previous chapter in which the word "or" in Bishop's "The Map" creates descriptive density within the ekphrastic network of speaker, map maker, printer, and reader—the syntactical significance of the word "or" destabilizes authorial control over the image and shifts the creative and interpretive relationship among them. I couldn't make that argument without focusing on the significance of a word that the stoplist preprocessing would remove from the data set altogether. This represents the crux of many debates over close and distant reading—losing something in order to gain another perspective.

Close and distant readings, however, are not mutually exclusive for LDA to be effective. By contrast, what the following tests proved to me is that they depend upon one another and, when used together, produce a richer understanding of texts—particularly ekphrastic ones. Unwilling to be convinced that removing high-frequency words could work, I designed a test to see how the presence of stopwords affected the usefulness of the topic keyword distributions the LDA produced. Without introducing the LDA process in detail here, as it is covered in greater detail in chapter 5, what is useful to know at this point is that LDA is an algorithm that sorts through documents and creates groupings of words that are most likely to co-occur—in other words, to appear in

the same document, in this case each poem. It is a form of machine learning that uses relationships between words to predict which documents share a common language.

As Mitchell, Heffernan, Hollander, Loizeaux and almost anyone writing about the genre explain, ekphrastic poems beseech their readers to "look" and "see," commenting to varying effect on the stillness and silence of the work of visual art. Therefore, the following test foregrounds the words: look, see, still, and stillness. To test the influence of stoplists on topic models of the poems in my digital corpus, I imported the dataset into MALLET in four ways and then ran a 40 topic model of each version of the preprocessed data. In the first test, I skipped preprocessing altogether, keeping every word in tact in the corpus (See Table 5). For the second test, I heavily edited the MALLET stoplist, and about 50% of the high-frequency words were removed from the poetry dataset before it was modeled (See Table 6, Appendix C). The third test only slightly modifies the stoplist leaving words frequently associated with ekphrasis in the text to be modeled (See Table 7, Appendix D). Finally, I ran the last test using MALLET's default stoplist (See Table 8, Appendix B). More attention in the following chapter will be paid to what, exactly, topics are and how they are created with LDA.[99] The following four tables demonstrate the results of the 40 topic key word distributions from each of the tests.

Learning to read topic keyword lists takes some practice; however, for the purposes of understanding the influence of stoplists on topic word distributions, there are a few things to focus on in the following Tables 1-4. First, each number on the far left represents the topic number, in this case from 0-39 because 40 topics were requested. The next number, called a hyperparameter estimation, shows the model's prediction as to

[99] A copy of the MALLET commands and parameters used for this test can be found in Appendix E.

how much of the collection might be described by each topic (For Topic 1, the hyperparameter is 0.25 or 25%). Next, to the right of the hyperparameter estimation are the top 20 words associated with the topic in descending order from most likely to least likely.

The question to ask as we compare these four tables is: how does the presence or absence of high-frequency stopwords in poetry data affect the distribution of words most likely to be found in a topic, and more specifically, how does it influence the distributions of ekphrastic texts? How might the presence or absence of words commonly associated with ekphrasis affect the ekphrastic poems within the model?

Table 5

Table 5 / Test 1: Keyword distribution - No stoplist		
Topic ID	Hyperparameter	Key word distribution
0	0.25555	the a in of on with house up by dog an table old street door at kitchen room under cat
1	0.83605	the and to s on at one from with this all out up now down back for time there no
2	0.00343	that ye and to in of for your my me so may ne thou is us doe sing woods which
3	0.00767	night moloch for wi johnny o ye auld york syne gat lang three fere andor stan owre kong lord fu
4	0.23175	we our us in are with when as how were for ourselves have what each who together from while re
5	0.50699	the of and in with that to out who on from god all fire time night world light earth man
6	0.38771	a and like with in as on skin his eyes hands hair black body inside mouth white blood little out
7	0.2099	of the its s this no to body death for from stone earth light here in into world by own
8	0.51036	was and a had were that in it to said but could did came when saw they then i one
9	0.68727	to not and be for will is love no but that if more or let all life heart as have
10	0.40124	the in of a for to on new from with old at or who one day year like years days
11	0.47055	t i it s you to can don but that what know so m say about a like ll we
12	0.05175	s little drink who mr dad money says for hair at richard boy shot spam black lamb milk get big
13	0.50871	i my me am in have when myself m so this mother see love face mine how man father own

14	0.08862	tree s fruit apple their sweet no come honey soil apples buy eat seed ripe garden leaf bees full bee
15	0.01868	de an la s n y t e a el le l me en miss green at din o on
16	0.03657	d o all soul n see ring heav th too r well while ev ever what thro long name pass
17	0.19007	a of s in blue white red as green with light gold flowers flower color yellow an glass eye silver
18	0.86073	and a to of that or it as for this so be not is have all what but if one
19	0.30406	a the of water like on with where river from into over through its or by across black an city
20	0.21519	they their them and are to men with by children women have up these those who themselves each old faces
21	0.65618	the in of like on light night sky wind as trees sun snow rain from moon leaves when at its
22	0.08392	the and to with of his in but their a on her nor as at this by so all thus
23	0.06979	one its of by s another life their when being each choose coat matter with movement hide person thread suddenly
24	0.04091	s too back whack jump off potato honey ice jazz ball happy chocolate crazy liver butter fast baby chicken bed
25	0.09293	poem a with write words poet poems by book an read or s writing poetry stop page word text written
26	0.23994	the and a with in s upon by their they blood from through its at like on heart each dead
27	0.13829	with thy thou that thee and o all shall from er nor yet but when now s then sweet which
28	0.43763	the to of it into then as and its up a down for on from over out back through head
29	0.24138	he his him man s and a for who father to at when dead on himself boy by now but
30	0.04181	no more than some any self nothing portrait even not duke less cause question change case necessary raven announce likely
31	0.27699	the of in to as an which by or with from more than on world mind life at sense these
32	0.36481	and the a of in s as are long old where that by all there song o so little is
33	0.04452	being not by people who because having states gertrude am inside american an been our also going during real plane
34	0.59288	the is a it in of that are there has not an this what but one nothing time will does
35	0.10043	the sea of and on ocean ship sand waves water boat tide as rock shore great beach shark land wind
36	0.24688	you your are in with have when will to who do at were yourself face re now woman ve for
37	0.18012	her she mother woman a s to as girl at from with for who eyes back white herself not hair
38	0.1903	the and to s of his in from by on who their god our land war great whose lord king
39	0.07415	a of each an horse letter line p em height box set space used ear ink horizon ocean between cowboy

Table 6

Table 6 / Test 2: Keyword distribution - heavily revised stoplist (stopword-ekphrasis-TT.txt)		
Topic ID	Hyperparameter	Key word distribution
0	0.02506	ball spam eat father milk potato eng trouble food from casey mrs horowitz chicken cow ate brooklyn ice never market
1	0.45871	had it of said in they then one did with out down saw would when for not on up went
2	0.22881	they their them are in by men children on have women up themselves see old hands faces girls heads bodies
3	0.19003	of with red white like hair black blue eyes girl mirror little lips green color yellow pink shoes brown silk
4	0.02321	gertrude has inside text stein lauren shot likely type bad everything must version species by animals across whitney effect genius
5	0.08197	poem write poems poet letter book read page writing poetry then words written wrote poets word name letters pages english
6	0.25807	he his him man in for on dead father himself boy has god now son eyes head hand see by
7	0.29525	in on with house room out for table up from street door morning window off into car kitchen outside night
8	0.37333	of in with for life death earth its from world soul god man no out light by dead where body
9	0.59167	of in on from where into light like over with down water by river up one then dark now here
10	0.16245	his from our by their on god in let men lord great every when for hell land good have own
11	0.01145	de la el miss le en me green do on yo ain thump verde les est no dat con ah
12	0.20948	of in with their for art self where praise his fire horse deep wall set double words blood broken rock
13	0.02364	black little richard moloch love daddy harlem white fat red high braids moonlight bloody bill jazz loves jesus european club
14	0.055	by form movement matter its stop coat desire theory point human consciousness physical solar above layer alive fabric tenure sensation
15	0.06074	thy thou thee love then for with art me thine ever doth er more sweet heart st dear soul from
16	0.00569	ye of ring with in let drink sing woods it which doe us for answer theyr eccho rats up out
17	0.04171	history here stone war from dead states fly name call exist names does mexico buzz monument marble built march between
18	0.41975	of in with for one old years when new on by two last year after how long from no first
19	0.12516	sea of on water boat blue ocean waves sand ship with land tide shore beach sun fish green island wave
20	0.10667	water with fish bones skin hot old smell out rock dirt broken bone sand some off mud steel dry dust
21	0.66363	will in are not love for one when no have day let night go on more may still time now
22	0.11132	our us ourselves from how together live bodies even heads return with heat occupation lives ours sleeping planet luminous hearts

23	0.09126	song music of bird singing sound sing songs notes long sings wind hear sang voice listen heard blow door sounds
24	0.58131	it don like know have say me how do when for about with they ll out on want up ve
25	0.08963	of in city america people over war new york under world white american man st streets street jew avenue paris
26	0.33956	in with like upon how eyes long night from by its voice heart they no light hand on hear deep
27	0.21709	of in tree green flowers grass flower fruit for trees garden like leaves sweet with summer apple rose wood leaf
28	0.34873	in of snow night sun sky on moon rain wind winter light summer fire are it white blue cold by
29	0.04999	of its bells hawk out from woman it guitar round itself stone time sand head parrot final shell living wings
30	0.00708	night wi sir conturbat mortis timor for auld lord some syne gat mr ye duke lang announce rocks owre cleft
31	0.05357	with his her he their on in from by more for heaven fate they stood mind now then vain arms
32	0.00422	for choose din ben hath ne instead thi erthe al gunga herte may merci alabanza oure yow myn shal no
33	0.04266	by new fear modern art times painting museum mr order model artist calm york studio public center situation prometheus oil
34	0.17635	her she of woman mother girl from for with hair in herself on not daughter child says one hands lady
35	0.5389	not of it in no for have which more by one would do much some will even how such us
36	0.62254	it like its in on into out up from back body down then over hand inside when hands their mouth
37	0.37389	my me in of have with mother love from would myself heart father mine life face hair name body god
38	0.4388	of in it are one like which on has nothing about something its from into world other time things sense
39	0.1928	of with from in their on now which when some where er by while her happy day round heaven still

When too many high-frequency words were left in the dataset, the signal-to-noise ratio becomes too high to interpret the word distributions, as can be seen in Tables 1 and 2. The results in Table 6 show the key word distributions when about half of the words from the MALLET default stoplist. The results produced by the model are heavily influenced by pronouns and prepositions. For example, Topic 2 forms around collective identities with key words including: they, their, them, are, in, by, men, children, on, have, women, up, themselves, see. Similarly, Topic 22 combines pronouns with collective

bodies, or many body parts, such as: our, us, ourselves, from, how, together, live, bodies, even, heads. While those may prove interesting if the study we hoped to perform were centered on collective versus individual identities or distributions between gendered pronouns; however, for the kinds of questions I want to focus on, strong pull along the lines of prepositions and pronouns is not as useful. For example in Table 6, Topic 33 includes the words "by, new, fear, modern, art, times, painting, museum, mr, order, model, artist, calm…." Despite the expectation set by the art-oriented vocabulary, the ekphrastic poems are not predicted by the model to include more than 4% of their language from that topic. However, the model inversely identifies that non-ekphrastic poems are highly likely to have a proportion of its language come from this topic, one more sign that the model is not producing useful results. Parsing the exact relationship between the documents that draw heavily from Topic 33 is not a matter of nuance. Instead the group seems to be primarily created around the use of the word "by." Having specifically included "by" in the model does seem to have made a difference in terms of locating and identifying poems in which "by" accompanies other kinds of words, many of which relate to other visual aesthetic objects, but sorting through the topic is about as useful as conducting any kind of close reading of the word "by" in a poetic collection. Consequently, the results are too disperse and don't help us to answer the questions we hoped to ask about "stillness" or "looking." Gendered pronouns such as those found in Topic 6 and Topic 34 (Table 6) might produce interesting studies about the use of gendered pronouns in poetry and the language that gendered pronouns tend to co-occur with; however, in this particular study, we're not simply looking at gender, but instead focusing on how women and men talk about stillness and looking. The way in which the

data is being sifted through the introduction of such high frequency words changes the focus of the model, and in this case, those changes are not productive to addressing the main questions of the project.

Returning to Table 5, the inclusion of all words in the topic model (ie. not using a stoplist at all) further reduces the potential uses of the model as a means for exploring the language of "stillness" and "looking;" instead, the topics generated in Table 5 demonstrate that articles, pronouns, prepositions, and high-frequency verbs are often found together. No surprises there.[100] Topic 8 is dominated by frequently used verbs: "was, and, a, had, were, that, in, it, to, said, but, could, did, came, when, saw, they, then, I, one." While the word "saw," which might be interesting in terms of understanding ekphrastic poetry, in this case it does little to identify a trend regarding ekphrastic "looking." Instead, by glancing down the hyperparameter estimations, one can find that the topics with the highest proportions, such as Topic 1, are distributions of articles and pronouns that offer little insight into the texts themselves. The key word distributions in Table 6 suggest that, contrary to my original inclination, even poetry uses enough high-frequency "little" words that their inclusion in the model only obscured whatever else might be discoverable.

[100] While the results of the third test are not useful for this particular study, the results could be useful for another line of inquiry. For example, Topic 25 in Table 2 seems exceptionally focused on poetry, writing, and words, accounting for an estimated 9% of the collection (.089 rounded up). To find a topic so clearly, semantically evident attests quantifiably to poetry's self-reflectiveness. What, then, is to be made of Topic 35, which draws heavily from the language of the sea (e.g. sea, ocean, ship, sand, waves, water, boat, tide, rock, shore, great, beach, shark, land, wind) and predicted to account for 10% of the corpus? Whether or not this is a bias of the dataset or a trend that deserves further inquiry is precisely the type of question that cannot be attended to here, but could be a trend worth investigating.

Table 7

Table 7 / Test 3: Keyword distribution - Slightly modified stoplist (stoplist2.txt)		
Topic ID	Hyperparameter	Key word distribution
0	0.0446	drink at cat goat wolf eat man fox hair dogs nose elephant milk shot black head cats ho play house
1	0.17536	dead blood death fire at war men black living man red god bones stone earth die burning broken hot iron
2	0.13577	sweet song er bird spirit heaven earth thoughts long dim deep ah heart hath soul beauty sad eye music weary
3	0.00975	bells de la el green en verde con los mi del se poem est thunder os ya poema oo sobre
4	0.01839	praise mr mrs johnny milk captain drawer horowitz good give yr friday henry kong learning nervous pigeon son alabanza run
5	0.10995	table dog kitchen eat ice coffee food bread plate orange morning cup dinner glass cat breakfast sugar cream fish happy
6	0.03259	year spring ring mary deer hills owl march joy pretty hall brings winter starlings wild turn months bells aged swarm
7	0.02432	parrot trumpet oil ringing kingdom beat lee surface chickens pink head parade strung poverty daybreak legged dust brain annabel lynn
8	0.03548	life earth choose world version flag creation announce rule pressure god animal persephone link witness discuss dream politics free space
9	0.21504	skin mouth body tongue blood bone heart inside flesh black face water hard fingers salt bones legs taste open cut
10	0.06537	god lord ye hell man good heaven soul truth holy jesus devil mercy peace sin prayer king son lamb spirit
11	0.02004	coat theory thread sir prometheus fabric sensation layer mattress matter completed fold folds yastrzemski cut threads stitch cloth code blue
12	0.21468	world mind made sense things at nature body thing point human space order form place part person thought feeling real
13	0.72897	night at light dark sun moon sky day wind sleep still stars rain long eyes hear air white darkness song
14	0.10302	land war freedom great power king america free still at er arms fame fight strong fate virtue battle man gods
15	0.62123	love life heart death time man world day soul long see god still earth men face eyes die before give
16	0.41795	body at back inside air world time small hands water after bird before eye look line tiny half place forward
17	0.22487	city house street window streets after windows room walking walls smoke past cars time train back houses car town door
18	0.05818	war american america people states richard white president free public big york great americans james bomb century flags bush plane
19	0.35331	white black woman hair red at girl eyes boy face hands blue back man mother bed room arms see mirror
20	0.11715	sea water sand ocean waves ship boat blue fish shore tide beach white green land sail island waters rock shark
21	0.00714	night wi fear monkeys auld syne gat fere lang fu owre ye na stack till afton nae lasses luve stac

22	0.03881	at place stood vain rest heaven high fear hell mind feast care fair prey length fate proud maid pride sight
23	0.00791	de miss le ain din la dat les cf slim ter dey jump ah sieve des scarlett pas yo peter
24	0.2945	tree green trees flowers leaves grass summer blue flower red fruit leaf garden white spring birds sun rose gold sweet
25	0.11396	see glass light eye words half surface portrait text picture dust hanging mirror box art black open lamp white oil
26	0.02932	man dead men woman women spam loves fish shoe age time european watch people world unfolded daddy sex jew lies
27	0.03353	horse see look passage perfect women beautiful cross enter miracle flow role alive beach real lower curious waits form interest
28	0.10424	poem write words poems letter poet book word read page writing poetry language letters wrote written days poets pages english
29	0.59076	at say don ll ve see time make says people day look things back good after feel remember thing won
30	0.00294	gertrude ye ne doe ring woods conturbat mortis timor theyr al eccho sing ben answer love stein thi shal erthe
31	0.12433	mother father children brother child home son parents made at years daughter school dead wife grandmother family days great dad
32	0.3327	at back saw before made turned looked thought head knew didn left still man stood night after heard felt sat
33	0.10409	stone head hand round eyes great man left rock foot stands hawk master back narrow set shape close mountain bound
34	0.03898	return sand life patience numbers lightning spiral shell rocks windmill weren animals live walrus desert call covered violet layer crab
35	0.02921	occupation whack time question cover west change grace duke wrong charity justice political chair circular mill fallen held diminish likeness
36	0.06473	thy thou thee art doth er hath thine sweet st ye happy tis joy fair mine hast view ere heaven
37	0.07777	men heard round day lay at saw till side fair merry stood high eyes good lady look sat young cried
38	0.18687	river water snow trees ice road cold lake winter fields country bridge sun stones horse woods field hill frozen leaves
39	0.01737	ball moloch york field father harlem times eng trouble casey jazz brooklyn tenure funeral thousand blues chang ebbets bat los

Table 8

Table 8 / Test 4: Keyword disribution - MALLET default stoplist		
Topic ID	Hyperparameter	Key word distribution
0	0.18396	round high bird moon blow full wide half silver low wild mountain wing green hills hill wind eye woods sun
1	0.02184	johnny good milk mrs run horowitz mr potato cream ate ice mattress pie learning plate barbie cow diner stan animal
2	0.12127	night moon stars light sleep sky star fly darkness dark air flight blue mother hour dusk birds midnight wings sight

3	0.08316	thy thou thee love er art sweet hath doth fair tis thine ye joy st heaven mine behold happy dear
4	0.21113	house window door room table dog kitchen morning bed glass line day windows floor back time sit wall half work
5	0.07156	didn people wasn weren worked knew war hide happened couldn family middle husband wanted occupation felt read age jews hadn
6	0.02495	horse moloch broken thump stone time west bit pony set rock angel cowboy greatest feet candle le mental jazz farewell
7	0.04728	man dead men woman cat dog loves fox dogs age wolf poor women caught desert clock friend tom lion walking
8	0.01104	de la gertrude el en green le din miss con yo verde inside dat ain les los slim dey ter
9	0.17388	black girl white hair woman red blue man shoes eyes big hat fat dress young girls back women wearing mirror
10	0.39171	heart love death soul eyes night long earth life day dead face sea heaven light sleep blood tears lips voice
11	0.64375	eyes back body face light hands hand head dark open air inside woman white world arms small feet close black
12	0.00981	choose mr life bo lady bonghy yonghy shalott sieve camelot jug order jones phyllis daphne tristan lands di heap jumblies
13	0.02399	women monkeys announce oil ferry political person boat advance waters whitman animals perfect great press flow beautiful bodies walt emperor
14	0.17747	don ll ve won back people hand good thing make didn drink bad left man love isn put wouldn kind
15	0.03142	war flags rise past lies captain passage great north rocks monument jew india thunder africa southern history country march british
16	0.02637	spam letter people american york henry president america war tenure world bush september william prometheus wallace army st guam yastrzemski
17	0.0023	ye ne sing doe woods thou conturbat mortis timor ring theyr al eccho ben answer hath love thi shal erthe
18	0.07099	music song bells sound sing singing notes praise time words songs guitar hear ear voice listen heard bird buy tune
19	0.78105	love time life day world things long years night make live find days back give home end place good call
20	0.2278	poem words book write word poems don read poetry poet feel work story page called makes writing letter language speak
21	0.18449	mother father boy children child son home brother years made dead girl school family sister daughter boys wife parents grandmother
22	0.01041	whack duke sir portrait cf text lord freud albert london grace water beat king jeoffry che ladies miss gri elizabeth
23	0.18265	sense mind world human body space life made things order form point nature place free movement time real history desire
24	0.18222	tree flowers flower summer green fruit sweet spring garden grass trees rose apple sun gold honey blossoms year autumn bloom
25	0.07575	water river lake surface ice back rivers ve bridge fish bottom swimming banks flat swim winter mississippi pan boat stream
26	0.01624	ball father blues field eng harlem trouble casey lauren boy shot white los play chang ebbets brooklyn yr people high

27	0.02369	fish riding drawer miracle moonlight woman fucking ray apologies kingdom opens daughter idea loved mermaid tlot thursday blades diminish pond
28	0.03332	fear maid mind place stood twas fair heaven death rest love ground pride care kind fate proud pursued prey pain
29	0.28902	skin blood mouth inside black tongue bone flesh eye water bones cut body glass eat broken salt hole open teeth
30	0.51021	wind sky light trees blue white snow leaves green rain sun field dark red grass air road earth water birds
31	0.02932	sand matter spiral stone living ancient edge endless fold cave silence rill burning canyon troy gate round slowly hush stretched
32	0.0315	states coat flag thread theory needle fabric gun cart dot version spider donkey united july call voice mexico moment rips
33	0.23052	god man great men earth world soul good make lord hell made peace death truth heaven fire spirit light time
34	0.0549	land ring er fame vain freedom fate arms heav war america race free plain liberty rise blood shore power sacred
35	0.07687	color stop painting art artist painted model perfect cover museum wrong makes reason wanted witness change completely case light hope
36	0.3379	back made head turned night looked left thought stood fell man heard knew put sat side long day hand men
37	0.15139	city street streets train past river inside cars paris car town york days houses talking night famous walking people lights
38	0.11498	sea water waves ocean ship boat sand tide fish shore beach rock great white shark blue land sail waters vast
39	0.00964	ye night wi ha auld merry tomlinson syne sin rats gat mayor piper michael man pipe lord alabanza fere lang

Perhaps the most demonstrative and telling difference between the results in Tables 3 and 4 is that in the topic where the words "look," "see," "still," "at," and "before" are included the distribution of ekphrastic poems become more diffuse.[101] Fewer ekphrastic poems are associated with any single topic in the third test (Table 7); whereas, in the fourth test (Table 8), the ekphrastic poems are more demonstratively clustered together. For example, in the third test, 88 ekphrastic poems were predicted to draw more than 1% of their language from Topic 13. Topic 13, likewise, is the most dominant topic across the collection and is associated with a predicted 72% of the poems overall, meaning the words with the greatest weight in Topic 13—"night, at, light, dark, sun, sky, day, wind,

[101] The exact list of words included in test 2 that were not included in test 1 can be found in Appendix B.

sleep, still"—are likely to be found in at least 72% of the corpus as a whole. Other topics from which many ekphrastic poems draw more than 1% of their language include topics 13, 19, 16, 15, and 29. Further sifting through of the poems, their genre classification, and their topic proportions reveals that the difference between "see" and "saw" or "look" and "looked" tends to decrease the coherence of ekphrastic poems. Contrary to my prediction, introducing the many-varied language of stillness and looking disrupted possible cohesion among ekphrastic poems and instead created affinities between texts that had more to do with the exact form of the word than with its semantic function.

Surprisingly, the fourth test (Table 8), which uses the MALLET default stoplist (removing the most words from the corpus before topic modeling), yields the most salient results. Topic 11, which describes 64% of the entire corpus but is not the most heavily weighted topic in the model, contains 125 ekphrastic poems—about 50% of the poems with the category tag, "ekphrastic." Topic 11's keywords, found in Table 8, coalesce around body parts (eyes, back, body, face, hands, hand, head, arms, feet), space (open, air, inside, world, small, close), and shade (light, dark, white, black). The poem most closely associated with Topic 11 is William Carlos Williams's poem "Danse Russe." Though we know Williams' poem was most likely written after having attended the Ballet Russe at the Met, the poem's close affiliation with a topic drawing heavily from the ekphrastic poems in the collection caused me to reconsider it. The Ballets Russes, which transformed 20th century ballet, synthesized efforts across the fine arts. Visual artists, including Pablo Picasso, Henri Matisse, Juan Gris, Giorgio de Chirico worked with Russian and French choreographers, producing sets, costumes, curtains, posters, and

even programs for performances.[102] Following the Ballets Russes at the Metropolitan Opera House, Williams' artist-friends at the 291 Gallery, were known to have painted and photographed Ballets Russes. There is no textual, or as far as I can tell critical, discussion of Williams' poem in terms of the visual arts; however, given Williams's prolific ekphrastic writing, his close relationship to visual artists at the 291 Gallery, and the shared language between "Danse Russe" and more than half the other ekphrastic poems in the collection, I am inclined to reconsider the poem as a form of ekphrasis, and as such it would serve as an interesting foil to those ekphrastic poems that take female bodies as their subject.

Most of the other ekphrastic poems included in this topic, however, draw a smaller proportion of their language overall from Topic 11, begging the question: from which other topics do the ekphrastic poems from Topic 11 draw their language? Moreover, how do those poems' topic distributions compare with poems tagged as non-ekphrastic? Prompted by the model, these questions appear most promising to the overall aim of the project. Using the MALLET default stoplist seems to sharpen the model's focus on the other discourses of ekphrasis as it has with representations of bodies, raising the possibility of exploring ekphrasis as drawing in varying proportions from multiple topics. Might ekphrastic poems that include language from Topic 11 also draw heavily from topics of love or mastery? Exploiting the metadata category for authors' gender, the results of the model introduce the possibility of exploring the distribution of topics among ekphrastic poems by men and by women and providing possible avenues to

[102] "Visual Art and the Ballets Russes." Ballet Russes Cultural Partnership. Boston University. Web. 16 Sept. 2012.

discover whether or not the distribution of topics within ekphrastic poems by men and by women reflect divergent attitudes toward the visual arts.

Closer examination of the model from Table 8 reveals that several of the higher probably topics are much less coherent. For example, Topic 19, a topic with which 66 of the ekphrastic poems in the data set are affiliated includes the words: love, time, life, day, world, things, long, years. At first, the relationship between these words appears vague; however, returning to the metadata and scanning the poems most closely associated with Topic 19, we find that the biblical verse Ecclesiastes 3:1 ("To everything there is a season,/ and a time to every purpose under the heaven.") draws most heavily from the topic's language distribution. The other poems also predicted by the model to draw heavily from Topic 19 share language that articulate the double-bind between love, time, and the physical constraints of the natural world as a limiting factor to human affection, for example: love that ends through physical death and separation, as in "To Dorothy" by Marvin Bell or spiritually as in "Psalm" by Alicia Ostriker, or emotionally as in John Dryden's "Why should a foolish marriage vow." More richly than I would have imagined possible, the algorithm picked out the subtle conflation of time and love and the language of ekphrasis, predicting that 66 ekphrastic poems drew from the language of love and time.

Furthermore, and significant to this study, Topic 35 in Table 8 predicts that 7% of the collection includes language that draws heavily from the visual arts: color, stop, painting, art, artist, painted, model, perfect, cover, museum, wrong makes, reason, wanted, witness, change, completely, case, light, hope. The model's prediction closely reflects our pre-existing knowledge from the metadata that approximately 5% of the database's poems are

ekphrastic. These findings are doubly relevant. First, by so closely estimating the number of poems that draw from the language of visual art the model promises a higher likelihood of identifying the distributions of language my study wants to explore. Second, by estimating slightly *more* texts draw from language closely allied to the visual arts, it offers the tantalizing possibility of discovery.

Conclusion

Ultimately, the stopword tests present a convincing case for using the MALLET default stoplist from the collection when running LDA topic models on the corpus of digitized poems described at the beginning of this chapter. Including words like "we" and "us"—pronouns and articles and commonly used verbs—in the topic model is not worth the additional noise that enters the results; however, I would argue that claims made about the topic should include a closer reading of the way excluded words affect a reading/understanding of the topic. For example, "still," "look," and "see" don't dramatically change the basic formation of the topics. The topics in Table 7 are very similar to topics in the Table 8. Searching for the terms in the topics doesn't do us much good, either. What is more interesting is to look at the topic distributions to see what other poems are more closely related to them and then to read the poems to see how the features we are trained to recognize compare in the poems also related to those topics.

As ekphrastic poems beseech their readers to "look" and to "see" more clearly, the ekphrastic poems themselves come into focus in the topic model better without actual words "look," "see," and "still" present in the dataset when it is modeled. In topics where ekphrastic poems are more evident, the words "see" and "look" are also commonly used terms; however, those same topics continue to form in models without the words at

all. They exist like ghosts in the data even without a physical presence. In fact, the topics in which the terms "see", "saw", "look" and "still" are most common in Table 7 happen to also be those topics that are most dominant in the model as a whole: topics 13, 14, 15, 16, 19, 25, 27, 29, and 32. While higher proportions of language from ekphrastic poems are likely to appear in those topics, the ekphrastic poems mirror similar trends among non-ekphrastic poems, too. In other words, the model shows that the ekphrastic poems in Table 7 tend to follow patterns that are detectable trends among most of the poems in the dataset. The topics from Table 5 and Table 6 are significantly less useful because of the signal to noise ratio. The topic keys create so much noise that the effort to understand them is not worth it. The question, then, is whether or not ekphrastic poems are "just like" other kinds of poems or if perhaps we should reread some of those poems to detect ekphrastic elements. Is it possible that more poems in the collection are ekphrastic than are tagged that way?

One might argue that close reading practices are susceptible to hyperfocus on high-frequency words. Concentration on terms such as "look," "see," and "still" helps us distinguish dominant trends in the genre, but the frequency with which those words are used tend to also skew their semantic context. Using the MALLET default stoplist during pre-processing and removing words frequently repeated throughout ekphrastic poems from the digital corpus foregrounds latent patterns of language that hint at the polyvocality and varied attitudes and discourses surrounding ekphrastic tropes that are often difficult to ignore such as the stillness of the image. As a possible example of Jerome McGann and Lisa Samuel's "deformance," LDA makes more obvious the differences between the "still unravish'd bride of quietness" in Keats' Ode from the

contemplative and restorative stillness of Carol Snow's "Positions of the Body." Noteworthy, too, is that LDA may suggest or predict these latent differences in discourses surrounding looking and stillness, prompting close readings and a return to the full, reassembled texts to consider them in relationships to one another, which represents the real hermeneutic potential for LDA studies of poetic corpra.

In the next chapter, I turn to the question of how to understand, use, and interpret LDA topics and reintroduce the network as the vehicle for reading the latent polyvocality of ekphrastic verse and the advantage this has for scholars interested in understanding the role ekphrasis by women plays in the genre's ecology. I return to the idea of topic distributions—the tendency for poems to draw from more than one of the same topics—as a way for understanding simultaneous discourses of ekphrasis, first within the collection of ekphrastic poems by themselves and later among the entire dataset. Through readings that begin with a macroscopic, network view and scale down into close readings and comparisons of texts, I demonstrate how the data collection, metadata creation, and preprocessing of texts from this chapter can be leveraged to develop a new methodology for understanding ekphrasis within its own tradition and in relationship to other poetic genres.