Casting Light on Invisible Cities: Computationally Engaging with Literary Criticism

Shufan Wang
University of Massachusetts, Amherst
shufanwang@cs.umass.edu

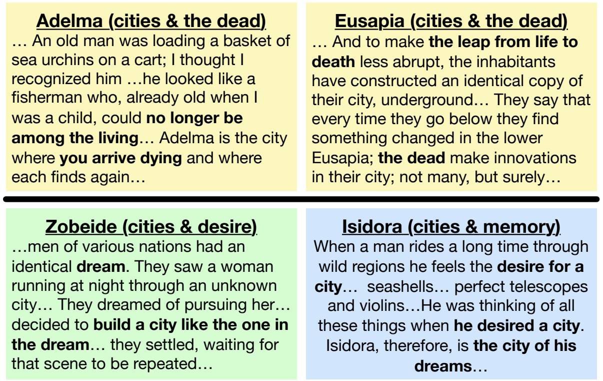
Mohit Iyyer
University of Massachusetts, Amherst
miyyer@cs.umass.edu

Abstract

Literary critics often attempt to uncover meaning in a single work of literature through careful reading and analysis. Applying natural language processing methods to aid in such literary analyses remains a challenge in digital humanities. While most previous work focuses on "distant reading" by algorithmically discovering high-level patterns from large collections of literary works, here we sharpen the focus of our methods to a single literary theory about Italo Calvino's postmodern novel *Invisible Cities*, which consists of 55 short descriptions of imaginary cities. Calvino has provided a classification of these cities into eleven thematic groups, but literary scholars disagree as to how trustworthy his categorization is. Due to the unique structure of this novel, we can computationally weigh in on this debate: we leverage pretrained contextualized representations to embed each city's description and use unsupervised methods to cluster these embeddings. Additionally, we compare results of our computational approach to similarity judgments generated by human readers. Our work is a first step towards incorporating natural language processing into literary criticism.

1 Introduction

Literary critics form interpretations of meaning in works of literature. Building computational models that can help form and test these interpretations is a fundamental goal of digital humanities research (Benzon and Hays, 1976). Within natural language processing, most previous work that engages with literature relies on "distant reading" (Jockers, 2013), which involves discovering high-level patterns from large collections of stories (Bamman et al., 2014; Chaturvedi et al., 2018). We depart from this trend by showing that computational techniques can also engage with literary criticism at a closer distance: concretely, we use recent advances in text representation learning to test a single literary theory about the novel *Invisible Cities* by Italo Calvino.



Adelma (cities & the dead)
… An old man was loading a basket of sea urchins on a cart; I thought I recognized him …he looked like a fisherman who, already old when I was a child, could **no longer be among the living**… Adelma is the city where **you arrive dying** and where each finds again…

Eusapia (cities & the dead)
… And to make **the leap from life to death** less abrupt, the inhabitants have constructed an identical copy of their city, underground… They say that every time they go below they find something changed in the lower Eusapia; **the dead** make innovations in their city; not many, but surely…

Zobeide (cities & desire)
…men of various nations had an identical **dream**. They saw a woman running at night through an unknown city… They dreamed of pursuing her… decided to **build a city like the one in the dream**… they settled, waiting for that scene to be repeated…

Isidora (cities & memory)
When a man rides a long time through wild regions he feels the **desire for a city**… seashells… perfect telescopes and violins…He was thinking of all these things when **he desired a city**. Isidora, therefore, is **the city of his dreams**…

Figure 1: Calvino labels the thematically-similar cities in the top row as *cities & the dead*. However, although the bottom two cities share a theme of desire, he assigns them to different groups.

Framed as a dialogue between the traveler Marco Polo and the emperor Kublai Khan, *Invisible Cities* consists of 55 prose poems, each of which describes an imaginary city. Calvino categorizes these cities into eleven thematic groups that deal with human emotions (e.g., desires, memories), general objects (eyes, sky, signs), and unusual properties (continuous, hidden, thin). Many critics argue that Calvino's labels are not meaningful, while others believe that there is a distinct thematic separation between the groups, including the author himself (Calvino, 2004). The unique structure of this novel — each city's description is short and self-contained (Figure 1) — allows us to computationally examine this debate.

As the book is too small to train any models, we leverage recent advances in large-scale language model-based representations (Peters et al., 2018a; Devlin et al., 2018) to compute a representation of each city. We feed these representa-

Proceedings of NAACL-HLT 2019, pages 1291–1297
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

tions into a clustering algorithm that produces exactly eleven clusters of five cities each and evaluate them against both Calvino's original labels and crowdsourced human judgments. While the overall correlation with Calvino's labels is low, both computers and humans can reliably identify some thematic groups associated with concrete objects.

While prior work has computationally analyzed a single book (Eve, 2019), our work goes beyond simple word frequency or n-gram counts by leveraging the power of pretrained language models to engage with literary criticism. Admittedly, our approach and evaluations are specific to *Invisible Cities*, but we believe that similar analyses of more conventionally-structured novels could become possible as text representation methods improve. We also highlight two challenges of applying computational methods to literary criticisms: (1) text representation methods are imperfect, especially when given writing as complex as Calvino's; and (2) evaluation is difficult because there is no consensus among literary critics on a single "correct" interpretation.

2 Literary analyses of *Invisible Cities*

Before describing our method and results, we first review critical opinions on both sides of whether Calvino's thematic groups meaningfully characterize his city descriptions.

The groups are meaningful: Some scholars believe that the thematic grouping imposed by Calvino reflects properties of the cities he describes; Vrbani (2012), for example, argues that Calvino's structure are "ontologically grounded in different ways". Buitendijk (2018) further provides examples of cities with the same label that are clearly thematically similar, pointing at the "cities of desire" as "informed by 20th century theories of desires associated with Sigmund Freud". Calvino (2004) himself claims that he creates most categorizations of cities with clear labels in mind, especially the cities of memory and desire, which he deemed as "fundamental cornerstones" of the novel. However, many critics argue that authorial intent is irrelevant when analyzing literature (Wimsatt and Beardsley, 1946; Barthes, 1994).

The groups are arbitrary: On the other hand, a large body of criticism focuses on the apparent mismatch between a city's assigned thematic group and the content of its descriptions. Bloom (2002) claims that the "cities are totally interchangeable"; Springer (1985) agrees, stating that "even the categories themselves seem both chosen and assigned arbitrarily". Teichert (1985) contends that "the catalogue is superimposed on, but does not cover, the elusive, fluid mass of an unwritten world".

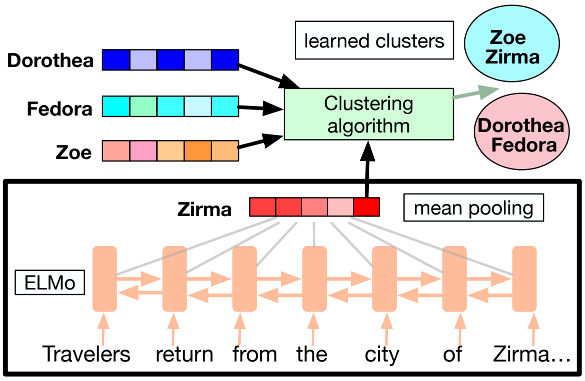


Figure 2: We first embed each city by averaging token representations derived from a pretrained model such as ELMo. Then, we feed the city embeddings to a clustering algorithm and analyze the learned clusters.

While out of scope for our computational analysis, many possible theories exist regarding *why* the groupings appear largely incoherent. For instance, Boeck (2004) posits that the structural incoherence exists because all of the cities actually describe different facets of Marco Polo's hometown of Venice. Breiner (1988) argues instead that Calvino's labels "may refer only to a projection of the Khan's occupational thirst for order, unrelated to the structure of the text", while Knowles (2015) hypothesizes that the mismatch is one of many obstacles that readers need to "untangle" to understand the central substance of the novel.

3 A Computational Analysis

We focus on measuring to what extent computers can recover Calvino's thematic groupings when given just raw text of the city descriptions. At a high level, our approach (Figure 2) involves (1) computing a vector representation for every city and (2) performing unsupervised clustering of these representations. The rest of this section describes both of these steps in more detail.

3.1 Embedding city descriptions

While each of the city descriptions is relatively short, Calvino's writing is filled with rare words, complex syntactic structures, and figurative language.[1] Capturing the essential components of each city in a single vector is thus not as simple as it is with more standard forms of text. Nevertheless, we hope that representations from language models trained over billions of words of text can extract some meaningful semantics from these descriptions. We experiment with three different pretrained representations: ELMo (Peters et al., 2018a), BERT (Devlin et al., 2018), and GloVe (Pennington et al., 2014). To produce a single city embedding, we compute the TF-IDF weighted element-wise mean of the token-level representations.[2] For all pretrained methods, we additionally reduce the dimensionality of the city embeddings to 40 using PCA for increased compatibility with our clustering algorithm.

3.2 Clustering city representations

Given 55 city representations, how do we group them into eleven clusters of five cities each? Initially, we experimented with a graph-based community detection algorithm that maximizes cluster modularity (Newman, 2006), but we found no simple way to constrain this method to produce a specific number of equally-sized clusters. The brute force approach of enumerating all possible cluster assignments is intractable given the large search space ($\frac{55!}{(5!)^{11}}$ possible assignments). We devise a simple clustering algorithm to approximate this process. First, we initialize with random cluster assignments and define "cluster strength" to be the relative difference between "intra-group" Euclidean distance and "inter-group" Euclidean distance.[3] Then, we iteratively propose random exchanges of memberships, only accepting these proposals when the cluster strength increases, until convergence. To evaluate the quality of the computationally-derived clusters against those of Calvino, we measure *cluster purity* (Manning et al., 2008):[4] given a set of predicted clusters M and ground-truth clusters D that both partition a set of N data points,

Method	Purity	Accuracy
Random	0.32	33.3
GloVe	0.35	35.9
BERT	0.40	39.3
ELMo	0.42	44.6
Human	-	48.8

Table 1: Results from cluster purity and accuracy on the "odd-one-out" task suggests that Calvino's thematic groups are not completely arbitrary.

$$\text{purity} = \frac{1}{N}\sum_{m\in M} max_{d\in D}|m \cap d|.$$

4 Evaluating clustering assignments

While the results from the above section allow us to compare our three computational methods against each other, we additionally collect human judgments to further ground our results. In this section, we first describe our human experiment before quantitatively analyzing our results.

Human clustering: We conduct a crowdsourced experiment to measure how well humans can disambiguate thematically different cities. Filling in the entire 55 × 55 adjacency matrix with human similarity judgments is expensive and time-consuming. Thus, we instead design a proxy "odd-one-out" task for collecting human judgments: given three city descriptions, two of which come from the same ground-truth thematic group and the other from a different group, workers are asked to identify the *intruder* city. We use the Figure Eight crowdsourcing platform[5] to collect three annotations each for 100 different city triples. Our interface initially displays only the first and last sentences of each city's description; workers can optionally click to reveal the full description. As workers are likely unfamiliar with *Invisible Cities* and its different thematic groups, this crowdsourced task provides a fair comparison to our computational approaches.

4.1 Quantitative comparison

We compare clusters computed on different representations using community purity; additionally, we compare these computational methods to humans by their accuracy on the odd-one-out task.

[1] The book contains a vocabulary of 5,372 word types, and the average length of a city description is 380 tokens.

[2] Using other composition functions such as the span representation of Peters et al. (2018b) had little impact on the learned clusters.

[3] The choice of distance metric (e.g., cosine, word mover) did not meaningfully impact our results.

[4] Purity ranges between 0 and 1, and a larger purity indicates a higher degree of agreement.

[5] Workers were restricted to English-speaking countries and paid $0.30 per judgment.

Purity of learned clusters: City representations computed using language model-based representation (ELMo and BERT) achieve significantly higher purity than a clustering induced from random representations, indicating that there is at least some meaningful coherence to Calvino's thematic groups (first row of Table 1). ELMo representations yield the highest purity among the three methods, which is surprising as BERT is a bigger model trained on data from books (among other domains). Both ELMo and BERT outperform GloVe, which intuitively makes sense because the latter do not model the order or structure of the words in each description.

Comparison to humans: While the purity of our methods is higher than that of a random clustering, it is still far below 1. To provide additional context to these results, we now switch to our "odd-one-out" task and compare directly to human performance. For each triplet of cities, we identify the intruder as the city with the maximum Euclidean distance from the other two. Interestingly, crowd workers achieve only slightly higher accuracy than ELMo city representations; their interannotator agreement is also low,[6] which indicates that close reading to analyze literary coherence between multiple texts is a difficult task, even for human annotators. Overall, results from both computational and human approaches suggests that the author-assigned labels are not entirely arbitrary, as we can reliably recover some of the thematic groups.

5 Examining the learned clusters

Our quantitative results suggest that while vector-based city representations capture some thematic similarities, there is much room for improvement. In this section, we first investigate whether the learned clusters provide evidence for any arguments put forth by literary critics on the novel. Then, we explore possible reasons that the learned clusters deviate from Calvino's.

Do learned clusters support existing analyses? The argument that *cities of desire* constitute a particularly coherent thematic group (Buitendijk, 2018) is partially supported by our clustering results. Three of the five *cities of desire* are grouped into the same cluster using BERT (two for ELMo), which makes it one of the most "internally coherent" groups. Similarly, some literary critics along with Calvino himself (Calvino, 2004) describe the *thin cities* as a fairly arbitrary group, which is supported by our results: when using BERT, no two *thin cities* are grouped into the same cluster. However, Calvino also suggests that the *cities of memory* group is a "fundamental substance" of the book and therefore should be highly coherent. Our computational methods cannot pick up this theme, instead scattering all *cities of memory* into different clusters.

Why do computers disagree with Calvino? In cases where the learned clusters deviate from the opinions of Calvino or literary critics, identifying the cause of the discrepancy is difficult: our computational methods are flawed, but there is also no one "correct" literary interpretation. Here we qualitatively analyze some of the learned clusters in an attempt to understand why the algorithm arrived at a particular assignment. First, we examine two cities from different thematic groups, *Beersheba* from "cities and the sky" and *Valdrada* from "cities and eyes", that belong to the same learned cluster (and are each other's nearest neighbors). The first two paragraphs of *Beersheba* describe a noble city "suspended in the heavens" with an identical but immoral "fecal" city underground, while the remaining paragraphs focus on the heavenly city. The description of *Valdrada*, which is built on a lake, shares this theme of twin cities: arriving travelers see "two cities: one erect above the lake, and the other reflected, upside down". While Calvino likely classified *Beersheba* based on its location in the sky, the two cities share undeniable thematic similarities. Rerunning the clustering algorithm after removing the first two paragraphs of *Beersheba* results in each city being assigned to a different cluster, which supports our hypothesis.

Another interesting case is the previously-mentioned "thin cities", supposedly bound together by airy and ambiguous themes (Knowles, 2015), which Calvino (2004) states were written after all of the other cities and are more incoherent than the other groups. While BERT does not group any thin cities together, ELMo categorizes *Isaura* and *Armilla* into the same learned cluster. The two cities appear largely dissimilar: *Isaura* is a city with a thousand wells dug by its inhabitants, while *Armilla* is an "unfinished" city without walls, ceilings, or floors. However, both cities' descriptions

[6] Fleiss $\kappa = 0.14$, indicating slight agreement, and two or more workers agreed on the intruder only 64% of the time.

mention supernatural beings living underground. In *Isaura*, some people believe "gods live in the depths" and "in the black lake that feeds the underground streams", while the last paragraph of *Armilla*'s description conjectures that it is "in the possession of nymphs and naiads" who "travel along underground veins". Removing these descriptions on underground gods and nymphs and rerunning our clustering algorithm yields a new assignment in which each of these cities belongs to different clusters.

When do humans and computers agree? Our computational approach yields generally comparable accuracies and more consistent results than human annotators in the "odd-one-out" task. On cities with concrete themes such as sky and trading, our approach with BERT and ELMo obtains accuracy of 0.44 and 0.45 respectively, (0.47 and 0.48 for humans). ELMo also performs on par with humans in some case: for example, humans achieve an accuracy of 42% on "cities and eyes", compared to ELMo's 43%. On groups where the theme word frequently occurs in the passage, such as "eyes", our approach even slightly outperforms the human readers. However, human readers are better at recognizing abstract intangible topics, such as memory.

6 Related work

Most previous work within the NLP community applies distant reading (Jockers, 2013) to large collections of books, focusing on modeling different aspects of narratives such as plots and event sequences (Chambers and Jurafsky, 2009; McIntyre and Lapata, 2010; Goyal et al., 2010; Eisenberg and Finlayson, 2017), characters (Bamman et al., 2014; Iyyer et al., 2016; Chaturvedi et al., 2016, 2017), and narrative similarity (Chaturvedi et al., 2018). In the same vein, researchers in computational literary analysis have combined statistical techniques and linguistics theories to perform quantitative analysis on large narrative texts (Michel et al., 2011; Franzosi, 2010; Underwood, 2016; Jockers and Kirilloff, 2016; Long and So, 2016), but these attempts largely rely on techniques such as word counting, topic modeling, and naive Bayes classifiers and are therefore not able to capture the meaning of sentences or paragraphs (Da, 2019). While these works discover general patterns from multiple literary works, we are the first to use cutting-edge NLP techniques to engage with specific literary criticism about a single narrative.

There has been other computational work that focuses on just a single book or a small number of books, much of it focused on network analysis: Agarwal et al. (2013) extract character social networks from *Alice in Wonderland*, while Elson et al. (2010) recover social networks from 19th century British novels. Wallace (2012) disentangles multiple narrative threads within the novel *Infinite Jest*, while Eve (2019) provides several automated statistical methods for close reading and test them on the award-winning novel *Cloud Atlas* (2004). Compared to this work, we push further on modeling the content of the narrative by leveraging pretrained language models.

7 Conclusion

Our work takes a first step towards computationally engaging with literary criticism on a single book using state-of-the-art text representation methods. While we demonstrate that NLP techniques can be used to support literary analyses and obtain new insights, they also have clear limitations (e.g., in understanding abstract themes). As text representation methods become more powerful, we hope that (1) computational tools will become useful for analyzing novels with more conventional structures, and (2) literary criticism will be used as a testbed for evaluating representations.

Acknowledgement

We thank the anonymous reviewers for their insightful comments. Additionally, we thank Nader Akoury, Garrett Bernstein, Chenghao Lv, Ari Kobren, Kalpesh Krishna, Saumya Lal, Tu Vu, Zhichao Yang, Mengxue Zhang and the UMass NLP group for suggestions that improved the paper's clarity, coverage of related work, and analysis experiments.

References

A. Agarwal, A. Kotalwar, and O. Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *In Proceedings of the Sixth International Joint Conference on Natural Language Processing,*.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *ACL*.

Roland Barthes. 1994. 11 the death of the author. *Media Texts, Authors and Readers: A Reader*, page 166.

William Benzon and David G. Hays. 1976. Computational linguistics and the humanist. *Computers and the Humanities*, 10(5):265–274.

Harold Bloom. 2002. *Bloom's Major Short Story Writers Italo Calvino*. Chelsea House Publishers.

Filip De Boeck. 2004. *Kinshasa: Tales of the Invisible City*. Leuven University Press.

Laurence Breiner. 1988. Italo calvino: The place of the emperor in "invisible cities". *Modern Fiction Studies*, 34(4):559–573.

Tomas Buitendijk. 2018. Port cities and desire in the work of italo calvino. *Port Towns and Urban Cultures*, (Article).

Italo Calvino. 2004. On "invisible cities. *Columbia: A Journal of Literature and Art*, (40):177–182.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *In Proceedings of the Joint Conference of ACL and AFNLP*.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *n Proceedings of the Thirty First AAAI Conference on Artificial Intelligence*.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,*.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have i heard this story before?: Identifying narrative similarity in movie remakes. In *NAACL-HLT*.

Nan Z. Da. 2019. The computational case against computational literary studies. *Critical Inquiry 45*, 23567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

J.D. Eisenberg and M. A. Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

D. Elson, N. Dames, and K. McKeown. 2010. Extracting social networks from literary fiction. In *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Martin Paul Eve. 2019. *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*. Library of Congress Cataloging-in-Publication Data. 9781503609365.

Robert Franzosi. 2010. *Quantitative Narrative Analysis*. Library of Congress Cataloging-in-Publication Data. SAGE Publication.

Amit Goyal, Ellen Riloff, and Hal Daume III. 2010. Automatically producing plot unit representations for narrative text. In *In Proceedings of Empirical Methods in Natural Language Processing*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jor dan L. Boyd-Graber, and Hal Daume III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *NAACL-HLT*.

Matthew Jockers and Gabi Kirilloff. 2016. Understanding gender and character agency in the nineteenth-century novel,. *Journal of Cultural Analytics*, culturalanalytics .org/2016/12/understanding-gender-and-character-agency-in-the-19th-century-novel/.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

Dominick Knowles. 2015. A redemption of meaning in three novels by italo calvino. *Digital Commons at Ursinus College*, English Honor Papers(2).

Hoyt Long and Richard Jean So. 2016. Literary pattern recognition: Modernism be- tween close reading and machine learning. *Critical Inquiry 42*, 23567.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *ACL*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Peter Norvig Dan Clanc and, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science (New York, N.Y.)*, 331(6014), 176182. doi:10.1126/science.1199644.

Mark Newman. 2006. Modularity and community structure in networks. *PNAS*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *North American Association for Computational Linguistics*.

Matthew E. Peters, Mark Neumann, Luke S. Zettlemoyer, and Wen tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *In Proceedings of Empirical Methods in Natural Language Processing*.

Carolyn Springer. 1985. Textual geography: The role of the reader in "invisible cities". *Modern Languages*, 15(4):289–299.

Evelyne Teichert. 1985. Words about nothing: Writing the ineffable in calvino and ma yuan. *Comparative Literature PhD Thesis, University of British Columbia*.

Ted Underwood. 2016. The life cycle of genres. *Journal of Cultural Analytics*, culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

Mario Vrbani. 2012. A dream of the perfect map calvinos invisible cities. *The Zone and Zones - Radical Spatiality in our Times*, (2).

Byron C Wallace. 2012. Multiple narrative disentanglement: Unraveling infinite jest. In *North American Association for Computational Linguistics*.

William K Wimsatt and Monroe C Beardsley. 1946. The intentional fallacy. *The Sewanee Review*, 54(3).