

# On the Automatic Tracing of Intertextuality by Meaning

**Walter J. Scheirer**

School of Engineering and Applied Sciences, Department of  
Molecular and Cellular Biology, Center for Brain Science  
Harvard University



# What is the connection?

Fie on't! ah fie! 'tis an unweeded  
garden,  
That grows to seed;  
things rank and gross in nature  
Possess it merely.

(Hamlet, Act 1, Scene 2,  
lines 135-137)

Thorns also and thistles shall it  
bring forth to thee;  
and thou shalt eat the herb of  
the field;

(Genesis 3:18)

# Similarity in meaning

“Thus, by judicious transfers and imitation he brought it about that, in what we read from another source in his work, we prefer his version or we marvel that it sounds better than its source. Let me say that he took from others from a half to about a full line of verse, then full passages slightly changed and rewritten, but with their meaning left intact, so that their source might be obvious...”

— Macrobius on Vergil, *Saturnalia*

# Intertextuality

- Kristeva: “Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another.”
- The Classicist: Two-word pairs are the most basic and common form of text reuse (*loci similes*).
- Hinds: A “poetic of corresponding inexactitude, which draws on but also distances itself from the rigidities of philological and intertextualist fundamentalisms alike.”

# Computational approaches

- Lee 2007, Global Linear Models
- Ganascia *et al.* 2013, Hash Coding
- Smith *et al.* 2013, Sequence Alignment
- Büchler *et al.* 2014, Noisy Channel Analysis (eTRACES)
- Coffee *et al.* 2014, Bi-gram matching (Tesseract Project)

# Limitation of lexical matching

Tesserae Benchmark:

Lucan's *Civil War* Book 1 vs. Virgil's *Aeneid*



A bust of Vergil in Naples 

Bi-gram matching  
necessarily misses  
**33%** of human  
identified parallels



Bust of the Roman poet Lucan 

# The process of reading

How did we identify the Shakespeare / Genesis parallel?

1. Examination of semantic content
  - Plant life → Gardens
2. Examination of sentiment
  - Something has gone wrong
3. Search of an internalized corpus
  - Eureka! **Garden of Eden**



# Semantic Analysis

- Popular strategy in natural language processing
  - Designed around the notion of word co-occurrence
    - ▶ Deerwester *et al.* 1990, **Latent Semantic Indexing**
    - ▶ Blei *et al.* 2003, Latent Dirichlet Allocation
  - Commonly deployed over large corpora
    - ▶ Jockers 2013, *Macroanalysis*



# Semantic analysis of small samples

We don't always have the luxury of a lot of text:

“Thy brother's blood the thirsty earth hath drunk.”  
(King Henry VI, Part 3.)

“And it came to pass, when they were in the field, that Cain rose up against Abel his brother”  
(Genesis 4:8.)

# Training set

1. And God said, Let there be light: and there was light.
2. And God said, Let there be lights in the firmament of the heaven
3. And God said, Behold, I have given you every herb bearing seed,
4. And on the seventh day God ended his work which he had made;
5. And the woman said, The serpent beguiled me, and I did eat.
6. And Adam called his wife's name Eve; because she was the mother of all living.
7. And it came to pass, when they were in the field, that Cain rose up against Abel his brother,
8. And Cain said unto the LORD, My punishment is greater than I can bear.
9. And Cain went out from the presence of the LORD, and dwelt in the land of Nod,
10. And to Seth, to him also there was born a son;

# LSI Matching




**Query: “Thy brother’s blood the thirsty earth hath drunk.”**

- 0. And God said, Let there be light: and there was light.
- 0. And God said, Let there be lights in the firmament of the heaven
- 0. And God said, Behold, I have given you every herb bearing seed,
- 0. And on the seventh day God ended his work which he had made;
- 0. And the woman said, The serpent beguiled me, and I did eat.
- 0. And Adam called his wife’s name Eve; because she was the mother of all living.
- 1. And it came to pass, when they were in the field, that Cain rose up against Abel his brother,
- 2. And Cain said unto the LORD, My punishment is greater than I can bear.
- 3. And Cain went out from the presence of the LORD, and dwelt in the land of Nod,
- 0. And to Seth, to him also there was born a son;

# Bag-of-Words model

7. And it came to pass, when they were in the field, that Cain rose up against Abel his brother
8. And Cain said unto the LORD, My punishment is greater than I can bear.
9. And Cain went out from the presence of the LORD, and dwelt in the land of Nod,

**Documents**

Terms	 7	 8	 9	
abel	1	0	0	
★ brother	1	0	0	← Only one shared word
cain	1	1	1	← Word Co-occurrence
field	1	0	0	
pass	1	0	0	
rose	1	0	0	
bear	0	1	0	
greater	0	1	0	
lord	0	1	1	← Word Co-occurrence
punishment	0	1	0	
dwelt	0	0	1	
land	0	0	1	
nod	0	0	1	
presence	0	0	1	

How did the Shakespearean phrase "Thy **brother's** blood the thirsty earth hath drunk" match these three bags?

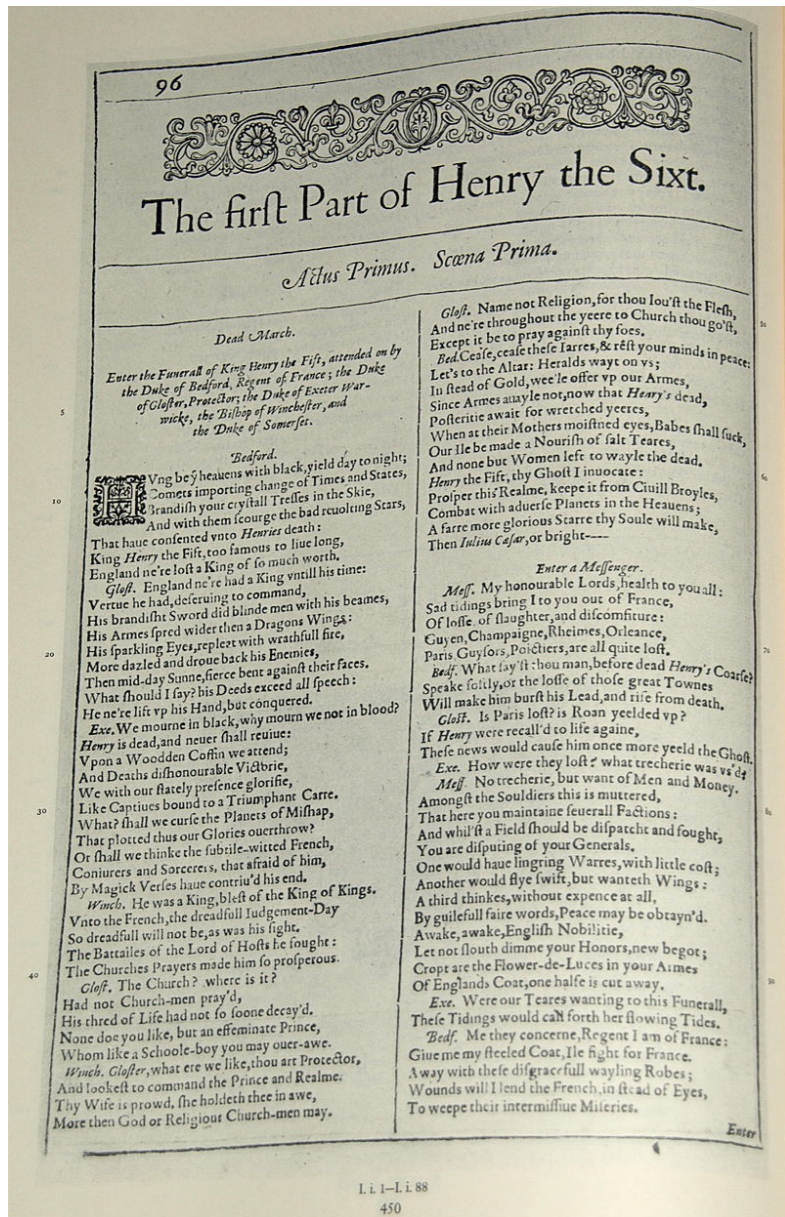
# LSI Algorithm: Training

1. Define a window to slide around the texts we want to match against (“documents”)
2. Apply stoplist to documents
3. Sample with window to collect a bag-of-words
4. Group bags into a document-term matrix
5. Reduce dimensionality with Singular Value Decomposition (SVD)
  - ▶ Free parameter: choose the number of “topics” to retain
  - ▶ This produces the model we’ll use for matching

# LSI Algorithm: Testing

1. Project a query in the model space, yielding a vector  $A$
2. For all of the document windows:
  - a. Project the window into the model space, yielding a vector  $B$
  - b. Compute cosine similarity between  $A$  and  $B$
3. Rank all of the similarity scores (higher is better)

# Why not try LDA?



“King”



“War”



“England”



“France”

Vector of Topics

# Why not try LDA?

“Thy brother's blood the thirsty earth hath drunk,  
Broach'd with the steely point of Clifford's lance;  
And, in the very pangs of death he cried,  
Like to a dismal danger...”

500 Character query; 1000 character documents

Gensim (Rehurek and Sojka, 2010) and MALLET (McCallum, 2002)

Run 1: 45:0.981, 1:0.311; 31:0.725  
Run 2: 66:0.455; 27:0.221; 45:0.156  
...  
Run  $n$ : 10:0.940, 27:0.931; 54:0.922



# Validation

- Small sample sizes
- Different languages
- Can we find new thematic matches?
- Can this tool help to produce new criticism?

# Case Studies

# Study #1: Latin Poetry

- Back to our Lucan / Vergil benchmark
  - *Civil War* Book 1: 685 hexameter lines
  - *Aeneid*: 9,896 hexameter lines
  - Parallels from four commentaries:
    - ▶ Heitland and Haskins (1887), Thompson and Bruère (1968), Viansino (1995), and Roche (2009).



TESSERAE

35 Missing Parallels  
(1/3 of total)

# Algorithm parameters

- Gensim (<https://radimrehurek.com/gensim>) LSI implementation
- Query window: 500 characters around and including a target line of text
- Document window: 1,000 characters around and including a line to match against
- Stoplist: 250 most common words
- Bag-of-words: inflected form of each word replaced with the set of all possible stems
- Number of topics: 10, 15, 20
- Count a hit in the top 50 results as a “match”

# Top match between a Query and Aeneid 3

qualis, cum turbidus Auster  
reppulit a Libycis immensum Syrtibus aequor  
fractaque veliferi sonuerunt pondera mali,  
desilit in fluctus deserta puppe magister  
nauitaque et nondum sparsa conpage carinae  
naufragium sibi quisque facit, sic urbe relicta  
in bellum fugitur. *nullum iam languidus aevo*  
*evaluit revocare parens coniunxve maritum*  
*fletibus, aut patrii, dubiae dum vota salutis*  
*conciperent, tenere lares; nec limine quisquam*  
haesit et extremo tunc forsitan urbis amatae  
plenus abit visu: ruit inrevocabile volgus.  
o faciles dare summa deos eademque tueri  
difficiles!

(Civil War 1.498 – 511)

postquam res Asiae Priamique evertere gentem  
immeritam visum superis, ceciditque superbum  
Ilium et omnis humo fumat Neptunia Troia,  
diversa exsilia et desertas quaerere terras  
auguriis agimur divum, classemque sub ipsa  
Antandro et Phrygiae molimur montibus Idae,  
incerti quo fata ferant, ubi sistere detur,  
contrahimusque viros. vix prima inceperat aestas  
et pater Anchises dare fatis vela iubebat,  
litora cum patriae lacrimans portusque relinquo  
et campos ubi Troia fuit. *feror exsul in altum*  
*cum sociis natoque penatibus et magnis dis.*

(Aeneid 3.1-12)

# Top match between Query and Aeneid 3

Just as when the swirling south wind drives the vast sea back from the Libyan Syrtes, and the shattered mass of the mast, with its sail, groans, the helmsman abandons the stern and leaps into the waves; and though the fittings of the hull are not yet strewn apart, each sailor fashions his own personal shipwreck; so too they desert the city and flee into war. Parents, frail with age, cannot call back their sons, nor wives, by their tears, their husbands; nor the ancestral homes, so long as they place their hopes on an unlikely salvation. No one hesitated on his threshold, to depart, perhaps, with a final look, filled with the love of his city. The crowd rushed on, heedless. How easily the gods give everything, how little they care to preserve it.

*(Civil War 1.498 – 511)*

After the gods saw fit to overturn the affairs of Asia and visit undeserved punishment on the race of Priam, after proud Ilium had fallen and all of Troy, built by Neptune, was a smoking ruin, we were driven by signs from the gods to seek exile far away and find vacant lands. Near Antander and the mountains of Phrygian Ida we constructed a fleet, though we were unsure where the fates were taking us, where we were to settle, and we gathered our men together. Summer had only just begun and my father Anchises ordered us to give sail for our destiny. I wept as I left the shores and harbors of my fatherland, and the plains where once was Troy. I was cast, an exile, onto the high seas, together with my companions, my son, the spirits of my household and the great gods above.

*(Aeneid 3.1-12)*

# Analysis

- Shared themes:
  - Abandonment (*diversa exsilia et desertas quaerere terras, litora cum patria lacrimans portusque relinquo*)
  - Naval imagery (*classem, vela, portus*)
- Roche: *BC* 1.504-7 and *AEN* Book 2; ***AEN 3.1-12***
  - Contrasting imagery (Aeneas's concern, Roman disregard)

# Bag-of-Words Control

- Does dimensionality reduction via SVD actually add anything?
  - Cosine similarity between bag-of-words representations
  - If LSI is better, we should see better match scores for relevant parallels



# Recovered Civil War 1 (BC) – Aeneid (AEN) commentator parallels

<b>BC Line</b>	<b>AEN Line</b>	<b>Shared Context</b>	<b>Topics</b>	<b>LSI Rank</b>	<b>LSI Prec.</b>	<b>BoW Rank</b>	<b>BoW Prec.</b>
1.60	1.291	Destiny of Caesar; peace	10	3	0.18	86	0.00
1.139	4.441	The blowing wind; tree	20	1	0.22	1	0.28
1.141	2.626	The blowing wind; tree	15	1	0.32	45	0.00
1.193	2.774	An apparition	20	33	0.08	47	0.00
1.193	3.47	An apparition	15	42	0.06	145	0.00
1.291	11.492	Horses	20	26	0.02	212	0.00
1.490	11.142	Flight	15	17	0.22	23	0.08
1.504	2.634	Abandonment	15	3*	0.52	70	0.00
1.504	3.11	Abandonment; Navy	15	4	0.16	215	0.00
1.673	2.199	Omens; terror	15	31	0.02	162	0.00
1.676	4.68	Dido as Bacchant	15	1	0.40	2	0.08
1.676	6.48	Prophecy	15	39	0.44	148	0.00
1.695	6.102	Frenzied discussion	20	22	0.20	31	0.20

\* also found by Tesserae lexical matching

Recovered 12 missing parallels

# Additional recovered Civil War 1 (BC) – Aeneid (AEN) commentator parallels with Bag-of-Words approach

<b>BC Line</b>	<b>AEN Line</b>	<b>Shared Context</b>	<b>BoW Rank</b>	<b>BoW Prec.</b>
1.1	4.628	War	48	0.02
1.8	12.313	Hostility	23	0.06
1.226	4.624	Broken treaty	13	0.10
1.226	12.435	Fortune	38	0.04
1.674	4.300	Dido as Bacchant	28	0.02
1.678	10.670	Questioning destination	17	0.16
1.685	2.554	Decapitation; shore	45	0.08

Recovered 7 additional missing parallels

# Exploratory work

<b>BC Line</b>	<b>AEN Line</b>	<b>Shared Context</b>	<b>Topics</b>	<b>LSI Rank</b>	<b>BoW Rank</b>
1.1	4.98	War	15	1*	1
1.141	2.252	The blowing wind	15	6	128
1.291	11.291	Conquest	20	1	4
1.353	1.647	City; nation	15	1*	26
1.504	3.639	Abandonment; nautical imagery	20	2	129
1.676	6.809	Bacchus	15	1	81
1.676	4.304	Bacchus	10	36	295

\* also found by Tesserae lexical matching

# Civil War 1.504 also matched:

sed fugite, o miseri, fugite atque ab litore funem  
rumpite.

nam qualis quantusque cavo Polyphemus in antro  
lanigeras claudit pecudes atque ubera pressat,  
centum alii curva haec habitant ad litora vulgo  
infandi Cyclopes et altis montibus errant.

*(Aeneid 3.639-44)*

But flee, you wretches, flee and slash the cables  
from the shore. For as great and tall as Polyphemus  
is who lives in his hollow cave, keeps wooly flocks,  
and milks their udders, a hundred such other  
monstrous Cyclopes live together along the curved  
shore, and wander the steep mountains.

# Bacchus re-contextualized

nam, qualis vertice Pindi

Edonis Ogygio decurrit plena Lyaeo . . .

*(Civil War 1.674-5)*

nec qui pampineis victor iuga flectit habenis

Liber, agens celso Nysae de vertice tigris

*(Aeneid 6.804-5)*

For just as a Thracian bacchant, filled  
with Theban Bacchus, rushes down from  
the summit of Mount Pindus . . .

Nor did Bacchus, who in victory guides his chariot  
with reins of vine, leading his tigers from the  
summit of lofty Nysa, [traverse as much land as  
Augustus will rule].

*And*

*BC 1.676 and AEN 4.300-3*

# Discussion

- Recovered 19/35 of the missing parallels
- What can we do better?
  - Algorithms that exploit context more like human readers
  - Revisit LDA? Need a solver that can handle small inputs

# Study #2: Fan fiction on the web

wattpad.com:

- 2 million writers
- 100,000 new pieces per day
- 20 million readers

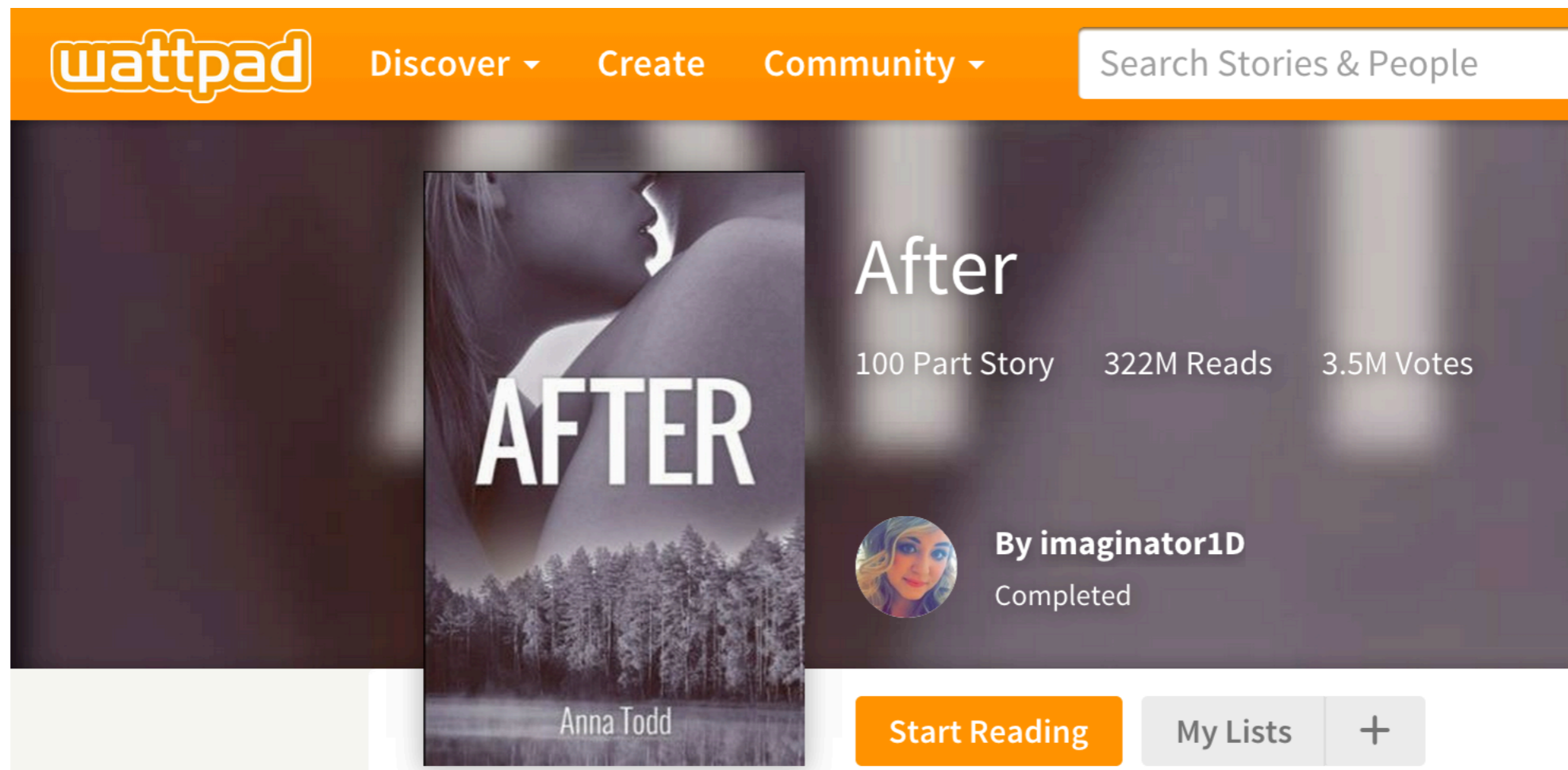


Image Credit: wattpad.com

# Song, dialogue, and the novel

“Some novelists draw on their own experience; others borrow from history, mythology or classic literary tropes. Anna Todd, a 25-year-old debut novelist in Texas, found inspiration in Harry Styles, the tousle-haired heartthrob from the British boy band One Direction.”

“Fantasizing on the Famous,” The New York Times, Oct. 21st, 2014





# Why fan fiction?

- Emerging literatures
- Direct participation in a popular narrative
- Interactive engagement with a text
- Which new texts will be successful?

**Publishers want to know**

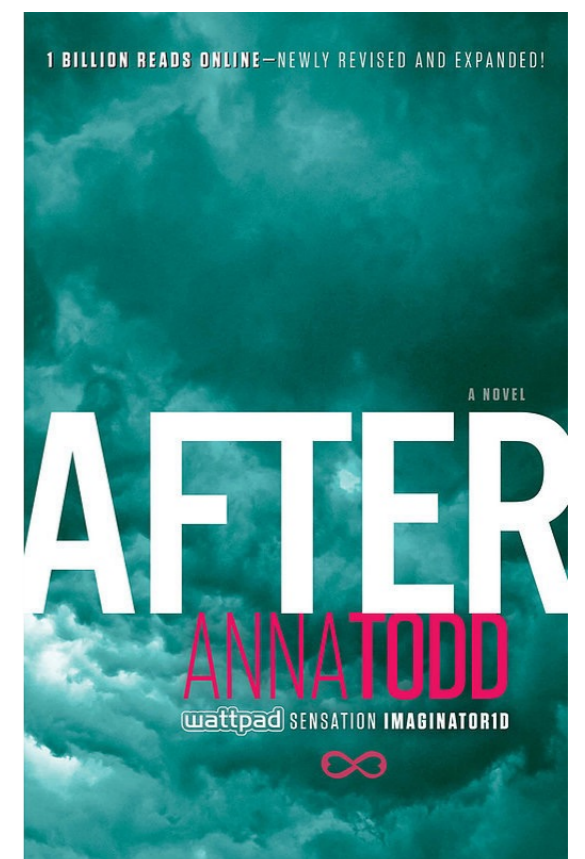


Image Credit: Gallery Books

# After

- Atypical style for a romance novel
  - Conversational first person narrative reminiscent of social media
- Obvious intertexts signaled via direct citation

Does the novel share something more with a broader literary culture?

[Recent Comments](#)

[Table of Contents](#)



**Itziar\_vr**

a day ago

This book is incredible I can't stop read it. I'm a fan of it!!  
Incredible Job :)



**NiallsFreakingPotato**

a day ago

this'll be my second time reading it!! so excited to read it again!



**namelessgella**

a day ago

Reading is fun the second time around lmao round 2 for this trilogy. Lez do dis

# Naïve Intertexts in *After*

“Today will be our last day on *Pride and Prejudice*, I hope you all have enjoyed it, for today’s discussion we will be talking about Austen’s use of foreshadowing. As a reader, did you expect her [Elizabeth] and Darcy to end up together in the end?” Professor asks and I raise my hand as always. Liam and I are usually the first to answer, and usually the only.

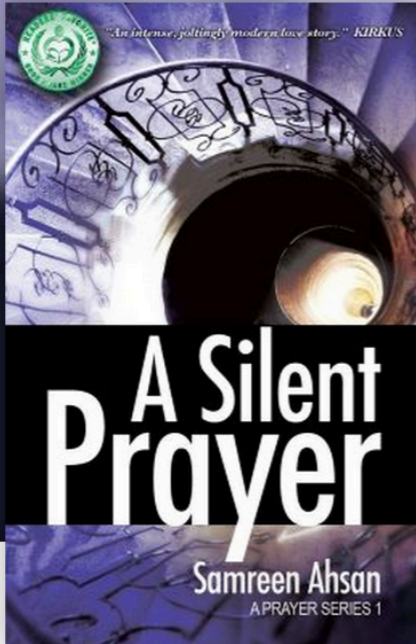
“Miss Young” he calls on me.

“The first time I read the novel I was on the edge of my seat to see whether they would end up together. Even now, as I have read it at least ten times I still feel anxious during the beginning of their relationship. Mr. Darcy is so cruel and says hateful things about Elizabeth and her family so I didn’t know if she would forgive him, let alone love him.” I answer and smile.

(“*After*,” Chpt. 24, Page 2)

# *A Silent Prayer*

- “Supernatural” Romance
- Also a statement on the first generation immigrant experience
  - Conflict between religious tradition and contemporary culture
- Clever use of allusion and a plot framed by the *Qu’ran*



**A Silent Prayer : A Prayer Series 1**

30 Part Story 235K Reads 2.7K Votes

By **SamreenAhsan**  
Ongoing - Updated 6 months ago

[Start Reading](#) [My Lists](#) [+](#) [Get notified](#)

# *A Silent Prayer* and the *Qu'ran*

“No. there are many, more than the human population. They have a world within our world, but we can't see them unless they want to be seen. There are males and females. There are good ones and bad ones, as well. You know what the interesting part is? Each human has been assigned a Jinni, a demon. We don't see it but it is always there. We recognize it when our soul is not strong enough to protect our body and our demon takes over, which makes us do all of the possible sins.”

(“A Silent Prayer,” Chpt. 16, Page 4)

Yet the foolish among us hath spoken that which is extremely false of GOD; but we verily thought that neither man nor genius would by any means have uttered a lie concerning GOD. And there are certain men who fly for refuge unto certain of the genii; but they increase their folly and transgression: and they also thought, as ye thought, that GOD would not raise any one to life. And we formerly attempted to pry into what was transacting in heaven; but we found the same filled with a strong guard of angels, and with flaming darts: and we sat on some of the seats thereof to hear the discourse of its inhabitants; but whoever listeneth now, findeth a flame laid in ambush for him, to guard the celestial confines.

(The Qu'ran, Chapter 72, 4-9)

# *A Silent Prayer* and the Bible

“Humans are made of clay. God collected seven different kinds of clay from the earth to create Adam. Clay supports and gives life, whereas, fire destroys and burns.” She paused for a moment and continues. “But like Satan, humans also carry pride, envy and arrogance, which sometimes work as a fire. There is a very thin line between humans and Jinn. Though they are invisible to humans, they do exist.

(“A Silent Prayer,” Chpt. 11, Page 1)

- Creation story is retold above
- The character of Adam shares his name with the biblical Adam
  - Emphasizes the genre’s trope of the primal man
- Adam’s sister is named “Eva”
  - Brother and sister are symbolic of temptation

# *A Silent Prayer* and Ismael ibn Kathir

Even after my rudeness, not calling him and thanking him for his kindness, he is still worried about me. After so many years, someone is worried about me. No one would believe that I've seen another side of Adam, which no one knows. *One who is not thankful to a person, is not thankful to God.* The prophet's quote buzzes in my mind and I decide I will call him to say thanks during my journey.

(“A Silent Prayer,” Chpt. 11, Page 1)

“This means respond to the poor with mercy and gentleness.” And proclaim the grace of your Lord. meaning, just as you were poor and needy, and Allah made you wealthy, then tell about Allah's favor upon you. Abu Dawud recorded from Abu Hurayrah that the Prophet said, Whoever is not thankful to the people, then he is not thankful to Allah.

(*Tafsir*, Book 93, 216-225)

# Corpus

- Primary works: complete texts of *After* and *A Silent Prayer* from [wattpad.com](http://wattpad.com)
- Comparison works:
  - The Book of Genesis (Authorized Version, digital text from Project Gutenberg)
  - The Qu'ran (translated by George Sale, digital text from Project Gutenberg)
  - *Tafsir* of ibn Kathir (digital text from the Internet Archive)



# Algorithm Parameters

- R “lsa” package
- Query window: 100 words
- Document window: 250 words
- Stoplist: 650 common words (Jockers 2014)
- Thresholds over cosine similarities

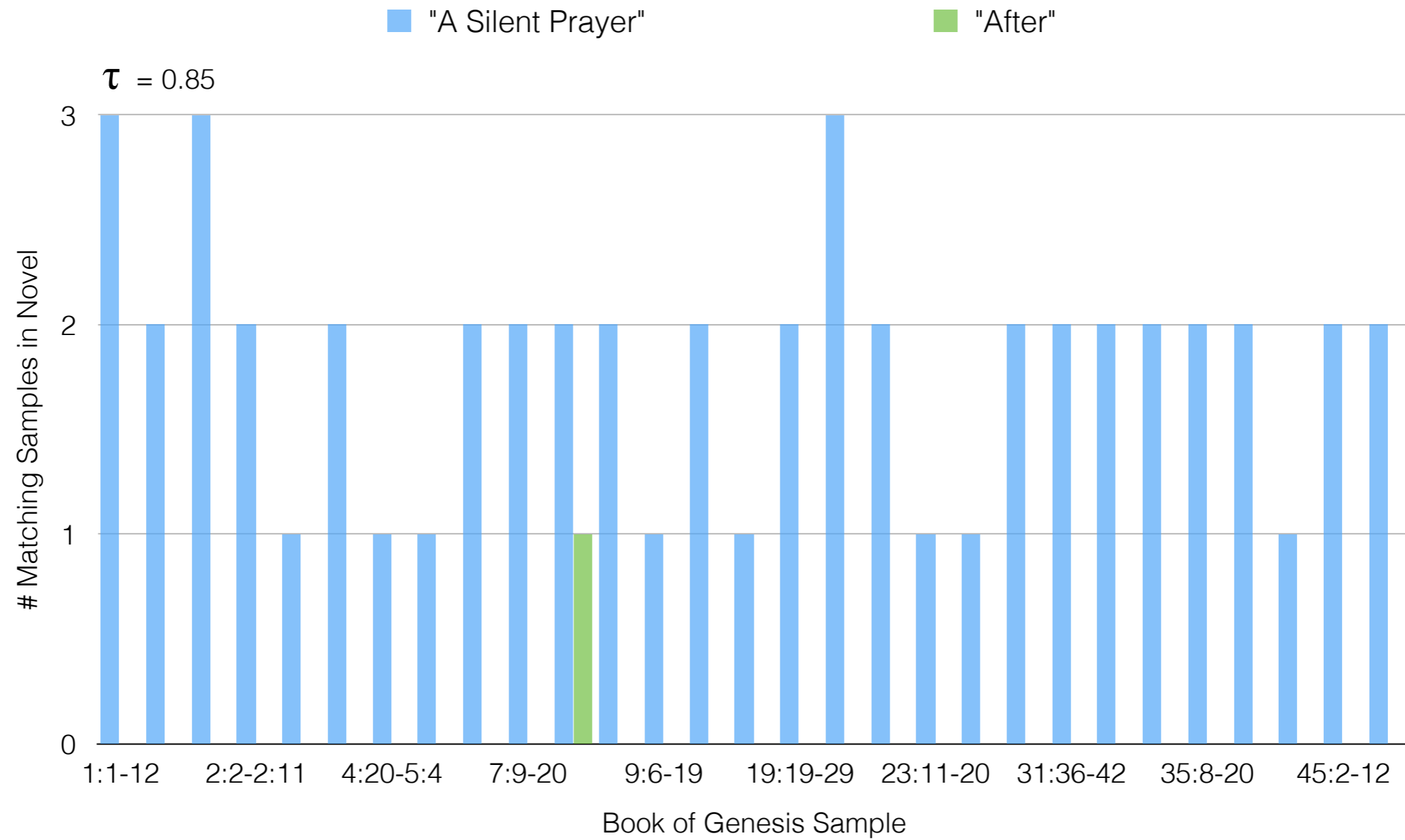
# Qu'ran Chapter 72 and *A Silent Prayer* Chapter 16

- Three query samples from Qu'ran Chpt. 72
- All samples from *A Silent Prayer*
- Threshold: 0.9; 20 topics
  - **Query 1: Chapter 16, Sample 4 (0.906)**
  - **Query 2: Chapter 16, Sample 3 (0.908)**      100% Precision
  - **Query 3: Chapter 16, Sample 3 (0.906)**
- Control: all samples from *After*
  - No matches

# *Tafsir* Chapter 93 and *A Silent Prayer* Chapter 11

- One query sample for *Tafsir*
- All samples from *A Silent Prayer*
- Threshold: 0.75; 10 topics 10% Precision
  - Chapter 6, Sample 4 (0.906)
  - Chapter 7, Sample 3 (0.785)
  - Chapter 7, Sample 6 (0.802)
  - Chapter 7, Sample 8 (0.772)
  - **Chapter 11, Sample 1 (0.753)**
  - Chapter 26, Sample 2 (0.817)
  - Chapter 26, Sample 6 (0.809)
  - Chapter 26, Sample 7 (0.766)
  - Chapter 28, Sample 3 (0.777)
  - Chapter 28, Sample 4 (0.795)

# Matches to Genesis



# Discussion

- *After* lacks literary merit compared to *A Silent Prayer*, but is massively popular
- Large scale automated study to understand readership trends
- How can we model an overarching relationship between a novel and its framing material?

Wrapping up...

# Resources

- W. Scheirer, C. Forstall, N. Coffee, “The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning,” to appear in *Digital Scholarship in the Humanities*
- Already in Tesserae
  - <http://tesserae.caset.buffalo.edu/cgi-bin/lisa.pl>
- R implementation will be on GitHub in the coming month

# Acknowledgements



Chris Forstall  
(U. Geneva)



Neil Coffee  
(U. Buffalo)



NATIONAL ENDOWMENT FOR THE

**Humanities**

Start-Up Grant Award #HD-51570-12



# Coming Soon...



## Quantitative Intertextuality

**Authors:**  
Walter Scheirer, Harvard University  
Christopher Forstall, University at Buffalo

A remarkable amount of information crosses our eyes and ears each day, yet we adeptly identify what is familiar with seemingly no effort at all. In many cases, what we see or hear has been shaped by recognizable prior sources. Such instances of *intertextuality* reveal a wealth of data about authorship, influence, and style, making them attractive targets for automatic identification. This volume introduces *quantitative intertextuality*, a new approach for the algorithmic study of information reuse in text, sound and images. Using a variety of tools drawn from machine learning, natural language processing, and computer vision, readers will learn to trace patterns of reuse across diverse sources for scholarly work and practical applications.

This is the first book to offer a rigorous, quantitative approach to the study of intertextuality. The authors are recognized experts in their respective fields, and have planted the seeds for this work by co-developing popular tools for textual and visual analysis. From practitioners specializing in forensics to students of cultural studies, readers from diverse backgrounds will find something of interest.

Questions?