Investigating the alignment of quantitative and qualitative literary analysis in American literature

Emma Angela Montecchiari

University of Trento
Computational Linguistics Course 2022/23
September 7, 2023

1 Motivation

The genesis of this project traces back to my academic foundation, where the beginning of the idea stemmed from my literary background. Specifically, it emerged from my fascination with a branch of literary criticism that originated in the 20th century within the Russian intellectual tradition. This movement, known as Formalism, sought to elevate the study of literature to an autonomous realm, enhancing literary criticism to the status of a discipline grounded in a scientific interpretation of artistic phenomena. The cruciality of this approach was a commitment to systemic and organic analysis, intricately within the formal, morphological, and syntactic attributes of literary works.

Moreover, building upon this paradigm shift towards more computational methodologies in critical analysis, emerged as an innovative approach in the late 20th century. This approach, driven by the desire to analyze a broader array of texts, aimed to enhance the conventional hermeneutic framework that had long confined literary criticism to a limited canon of works. The former approach, often termed "close reading" in the scholarly discourse, remained rooted in analytical and qualitative exploration. In contrast, the newer approach, referred to as "distant reading," embraced a synthetic and quantitative perspective, propelling literary analysis into new dimensions by scrutinizing texts en masse, thus enabling a deeper understanding of literary phenomena on a grander scale.

2 Literature

2.1 Close vs. Distant Reading

While the close reading of a text is the traditional method in literary criticism that was broadly used and developed mainly in the middle of the 20th century, distant reading is a rather novel idea that was mainly introduced by Franco Moretti at the beginning of the 21th century. Close reading is the fundamental method in literary criticism. Nancy Boyles (Boyles, 2013) defines it as follows: *"Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension."* In other words, close reading is the thorough interpretation of a text passage by the determination of central themes and the analysis of their development. Moreover, close reading includes the analysis of individuals, events, and ideas, their development. While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s). Moretti (Moretti, 2000) describes distant reading purposes as allowing *'to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small or the very large, the text itself disappears, well it is one of those cases when one can justifieably say, less is more'.*

The new approach does come with some trade-offs. As Moretti continues in the mentioned citation (Moretti, 2000): *'If we want to understand the system in its entirely, we must accept loosing something'.* Despite its potential and innovation, it faces significant issues related to methodology and its subject matter. While it allows for the collection of vast amounts of data, the technical methods often struggle to construct analyses as intricate and precise as those produced by traditional hermeneutic approaches. It sacrifices the finer details, subtle meanings, and nuances like irony that traditional analysis excels in capturing.

However, it aims to determine whether considering broader literary patterns from an empirical standpoint can produce valuable insights that can complement traditional analytical readings. This challenge is amplified because of the inherent complexity of the subject matter itself—an artwork intricately crafted from linguistic elements, rich in subtleties meant to convey both individual and collective expressions through its unique features. A melting of the two then is suggested from Jockers (Jockers, 2013): *'Micro-oriented approaches to literature, highly interpretative readings of literature, remain fundamentally important. Just as microeconomics offers important perspectives on the economy. It is the exact interplay between macro and micro scale that promise a new, enhanced and perhaps even better understanding of the literary record. The two approaches work in tandem and inform each other. Human interpretation of the "data", whether it be mined at macro or micro level, remains essential. While methods of enquiry, of evidence gathering, are different, they are not antithetical, and they share the same ultimate goal of informing our understanding of the literary record'.*

2.2 Distant Reading Techniques

In the realm of formal text analysis, my focus diverges from traditional stylistic analysis as it zeroes semantic implications within the computational sphere, primarily concerning surface-level textual phenomena tied to correspondences and high-frequency terms. While my initial intention was to create word and document embeddings or engage in topic modeling, due to limited textual resources and the potential imprecision of fine-tuning or applying pre-trained models, I opted to work exclusively focusing on the available texts.

In order to gain this type of understanding of the texts, I applied a range of data mining metrics and techniques (Rockwell et al., 2022). Specifically, I utilized readability indices such as the **Flesch Reading Ease** and **Gunning Fog Index**. These indices are designed to evaluate the readability of a text by considering factors such as the length of sentences and the number of syllables in words (Van de Rul, 2023). They help determine how easily a text can be comprehended by its audience. Additionally, I used other readability tests like the **Dale-Chall formula** and the **Automated Readability Index**, which provide further insights into the text's accessibility and readability. Furthermore, to analyze the variability within the texts, I employed metrics like the **Hapax Legomena Ratio** and the **Type-Token Ratio**. The hapax legomena ratio identifies the proportion of words in the text that occur only once in relation to the total number of words, providing a measure of word uniqueness and diversity. On the other hand, the type-token ratio investigates lexical richness by examining the relationship between the specific words used (types) and their overall frequency (tokens). These measures offer valuable insights into the text's vocabulary diversity and complexity.

Among stylistic indicators, my focus has been also towards identifying frequently used terms as expressions of the author's idiolect, i.e. authorial imprinting among which is the repeated use of favorite words defined as high frequency lexicon. In this sense, the division between grammatical words and lexical words is useful in defining two different speech categories. The former express grammatical relationships between words that make up sentences, the latter constitute elements of utterance with higher semantic content by configuring themselves as more variable and substitutable. Keeping into account text length and form, in order to calculate the term frequency I employed the **TF-IDF** (Term Frequency-Inverse Document Frequency) metric. This technique assigns weights to terms based on their importance within individual documents relative to the entire corpus, generating TF-IDF vectors, as a set of weighted terms. In order to compute some similarity analyses I used **Cosine Similarity** using TF-IDF vectorization. When applied to TF-IDF vectors, cosine similarity quantifies how closely related two documents are in terms of their content. Documents with similar topics or themes will have higher cosine similarity scores, making it possibly as a valuable tool for tasks like movements typicality.

3 Research Question

My aim has been to employ quantitative computational language analyses to characterize throughout literary genres. I wondered if this quantitative approach could somehow intersect with traditional genre categorization. Could similarities found in the repetition of terms and forms describe genres similarities, differences, and characteristics? Could a description of co-occurrences and relationships shed light on aspects derived from a qualitative analysis perspective?

These questions were the driving force behind this project, which seeks to provide and complement a description of an object that defies narrow classification. It aspires to complement a more meticulous and sophisticated analysis. The question that remains is whether this interpretive quantitative surplus value can enhance the interplay between textual subjectivity and the interpretations and definitions that surround it.

4 Proposal

I am investigating whether quantitative investigative techniques align with qualitative analysis, exploring the consistency between traditionally attributed characteristics of literary movements and computational methods. This project aims to test if distant reading analysis can complement and validate close reading techniques, exploring the potential synergy between the two. To achieve this, I focused on a diverse corpus of literary movements, selecting one for in-depth analysis. I employed stylistic characterization techniques within the texts to describe these movements and assess their alignment with traditional categorizations. Subsequently, I conducted a more refined analysis of a specific genre and compared it to traditionally assigned characteristics. Computational linguistics techniques were then applied to identify similarities and differences among movements, broadening the scope of analysis. Throughout this research, I incorporated external material from classical qualitative analyses of the chosen movements, allowing for a dual perspective while predominantly applying quantitative methods.

The genre I chose for a detailed comparison was American Gothic. It stood out due to its pronounced stylistic elements, featuring allusive imagery and less common themes. Additionally, it offered the advantage of a wealth of critical materials for reference. While the genre's definition may exhibit some instability and polysemy, certain recurring characteristics, such as vivid descriptions of places (like castles and underground passages) and a consistent aim to evoke terror through elements of mystery, cruelty, and horror, form the core of its characterization (Perazzini, 2013). These characteristics serve as the foundation upon which interpretations of works within the genre are constructed. This inherent complexity and the genre's elusiveness make it a valuable candidate for analysis with computational tools, as it can potentially reveal connections and differences to the widespread features found in literature.

5 Project

The source code and the scripts used in the project, as well as the corpus data and the graphs of the results, are inside a repository on GitHub that you can find at this link: https://github.com/memonji/Computational_Linguistics_2022-23.git.

5.1 Data Collection

On the initial phase of my research, I focused on building a collection of literary genres. This involved manual searches and selections to create the specific genres' corpora. The decision to concentrate on 19th- and 20th-century American movements was driven by the practicality of accessing texts on online open-source platforms and utilizing established language analysis tools for English language processing. I organized texts based on traditional literary history classifications, selecting works from classic authors within their respective genres. These texts were sourced from Project Gutenberg. I cleaned the texts by removing standard Gutenberg format headers and footers as necessary.

Within this corpus, I prioritized canonical authors representing their respective genres, forming the ground of my database as key figures in the literary canon. While the texts I selected were widely accessible, I acknowledge that they may not fully capture the richness and diversity of a comprehensive corpus, which could include more informative and structured materials for a richer and more grounded understanding.

5.2 Complexity measures spread over movements

The analyses I performed has been developed mainly with Python and Jupyter Notebook. Firstly, I computed complexity measures within different filters for the texts, to derive a series of features related to statistical measurement.

In particular I did a measure of *average phrasal length*, ratio of words in the text to its sentences. Then two interrelated metrics such as the *average syllable number* within words and the *presence of words with more than two syllables*, which I connected with difficulty in reading the text. Then I computed different metrics of text reading difficulty, so that I was not chained on one computation. Specifically *Flesch Reading Ease, Gunning Fog, Smog Index, Dale Chall Readability Score, Automated Readability Index.* As lexical variability I also connected my analysis to the computation of hapax in relation to text length through the *Hapax Legomena Ratio* metric. Overall lexical density was also computed through the *Type-Token Ratio.*

I performed this analysis for all the selected texts of the literary movements, then computed the movement average for each index. For better visualization, I plotted bar graphs.

5.3 Most frequent words on Gothic

As for the close reading more specific analysis I focused on the analysis of top frequency lexical words for the American Gothic genre. I first considered computing words for the entire movement without distinction by grammatical category. Since the results were not very telling, however, I decided to focus on categorical differentiation that would allow me to stylistically differentiate and better analyze the results. For this purpose I constructed four features for the different grammatical categories keeping the texts with the presence of stopwords, lemmatizing and tokenizing them. The close and analytical analysis was done through the analysis of terms drawn from the most frequent words through the TF-IDF filter. I derived different frequency lists (with a window of 100 terms), particularly noun, adverbial, adjectival, and verbal classes. In relation to the different features of these different classes, I carried out an analysis differentiating them according to their semantic field. After then, I associated these retrievals with the characteristics typically associated to Gothic by close reading analyses (such as for instance sublime features and medieval settings) Text vectorization was done through TF-IDF matrices and subsequent extraction of features that matched the grammatical class. I performed this analysis for the chosen movement to undergo the close reading analysis, Gothic. After then, for Transcendentalism, for which I put the results in Appendix 1, in order to have a comparison counterevidence (Cannella, 2015).

5.4 Cosine Similarity between movements

The last analysis was that of cosine similarity, inner to the movements themselves and between them. This, too, was done through construction of vectorization on the basis of TF-IDF after stopwords cleaning, lemmatization and tokenization. The cosine similarity was constructed on the individual matrices (corresponding to the individual corpora), to test the inner similarity of the different movements, and between them. I used the results, in matrix form, in two ways. First by plotting some heatmap graphs that I considered useful for the analysis. Secondly by plotting the result averages for the different matrices on a plot bar graphs.

6 Discussion

6.1 Top Frequency Words Gothic

The overall results are in Appendix 2. They are the retrieval of the top frequency TF-IDF words in a window of 100 terms. I am performing the analysis over the verbs, nouns, adverbs and adjectives throughout the semantic class in which I classified them.

6.1.1 Verbal Class

As for the verbal class, we observe four predominant semantic categories. The first revolves around motion actions, like *return, continue, bring, pass, reach, proceed, walk, discover*, constituting perhaps the most numerous class, characterized by progression and exploration. This mirrors the significant role of movement within the Gothic genre, where characters often embark on adventurous journeys, in search of an object of desire or in the restlessness of their psychological condition. In contrast, actions of stasis correspond to more perceptual moments, like *believe, observe, remember, feel.* This category is closely linked to reflexive and psychological states, reflecting the Gothic's emphasis on character introspection and psychological complexities, often marked by insecurity, instability, ambiguity, and introspection about one's state. This category also seems to be very much connected to the imaginative state toward the future. Indeed, we have a third semantic category that we can cross that corresponds to that of adventure. We find verbal actions related to *winning* or *losing, falling, entering* for example. These actions commonly depict encounters, confrontations, and moments of self-discovery typical of the Gothic genre. Additionally, life and death, intrinsic elements of Gothic fiction, are recurring images within this category. Lastly, we find a semantic category that captures relational and dialogic instances, like *tell, speak, ask, reply*, reflecting the abundant presence of dialogues within the texts.

6.1.2 Nominal Class

Within the noun class, we observe several key categories. Firstly, there are spatial descriptors pertaining to the environment, including *length, manner*, and *sound*, as well as temporal elements like *night, day*, and *time.* Additionally, nouns related to spatial positioning, such as *country, door, window,* and *wall*, feature prominently. Some of these spatial elements are associated with the medieval setting, typical of Gothic, such as *abbey, chamber*, and *village.* Another category includes descriptive nouns linked to the sublime, encompassing terms like *night, dream, spirit*, and *soul*, often intertwined with natural references. Descriptors of the body are then found, which is in fact used and described to emphasize the corporeality of the characters' experiences. Nouns characterizing the body are present, emphasizing the corporeal aspects of characters' experiences. In addition, there are nouns related to the social context, both vertical and horizontal, including structural-political terms like *law, king*, and *power*, as well as those pertaining to characters' social relationships during their adventures, such as *lady, gentleman*, and *friend.* Lastly, meta-narrative elements are evident, highlighting the presence of the writing instance, with nouns like *poem, character*, and *illustration* underlining the narrative dimension of the texts.

6.1.3 Adjectival Class

In the adjectival class, two primary categories stand out: descriptive adjectives and those related to environments and places. The first category is closely tied to the realm of the sublime, featuring descriptions of elements that evoke enormity, atypicality, monstrosity, ambiguity, and melancholy. This category is rich in adjectives that convey *abnormal*ity and *sad*ness, along with descriptors of the *wild, vast*, and *ancient.* In contrast, the second category emphasizes otherness through adjectives associated with the psychological states experienced by characters and the situations they encountered. Positive traits and neutrality are found here as well. Similar to the verbal class, which described movement, this category focuses on describing places, often intricately connected to the atmospheric settings prevalent in Gothic literature. Interestingly, it's worth noting the presence of links to historical instances, characterizing the typical medieval setting of the novels.

6.1.4 Adverbial Class

In the adverbial class, we encounter several notable categories. Firstly, there is a significant presence of adverbs related to time and motion measurement, reflecting elements of movement, character investigation, and perception. Such as *immediately, suddenly, slowly, somehow* underlying some suspense stylistic elements. Another substantial category comprises typical discourse expressions, which add nuance to the dialogue and interactions within the texts. Furthermore, we find adverbs associated with measurement, like *scarcely, gradually, nearly*, possibly reflecting descriptions of places and journeys. Within this category, there are also adverbs with sublime connotations, such as *deeply, extremely*, and *exceedingly*, further contributing to the text's overall atmosphere and intensity.

6.2 Complexity measures

The data mining analyses came to indicate indices for the different statistical filters which I previously described, comparing texts within each literary movement. Observing the average results for the movements in comparison to the Gothic one, we can talk about some results. Known for its distinctiveness and strong characterization, we observed atypical long phrasal structures compared to other literary movements (Fig 1).

Similar and on average compared to other literary movements, on the other hand, is the index of lexical richness. It aligns with the widespread influence of this genre and its canonization, often revisiting established themes from its English Gothic precursor. The Gothic genre is also exhibiting low in respect to the use of hapax (Fig 2) compared to the general average, especially when contrasted with Dark Romanticism, which displays a distinctive inventiveness. The readability of the Gothic works, which I produced through the different metrics is almost high for all the indices within the corpus set of documents. This, along with the word difficulty index, underscores the genre's descriptive nature and its dense exploration of settings and character dynamics within various passages. The graphs of the results are in Appendix 2.

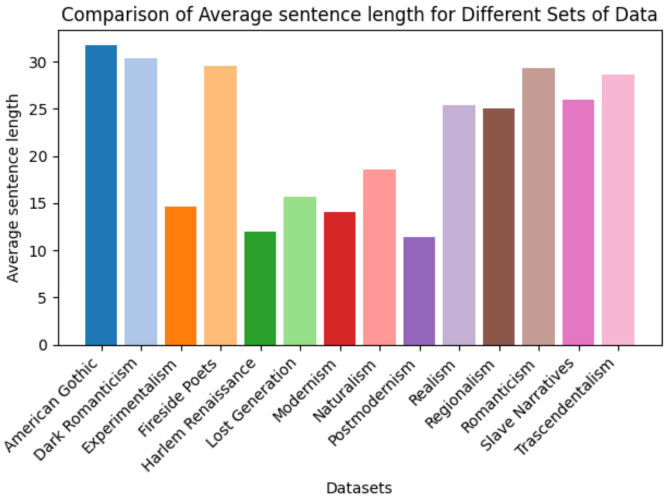

Figure 1: Average Sentence Lenght

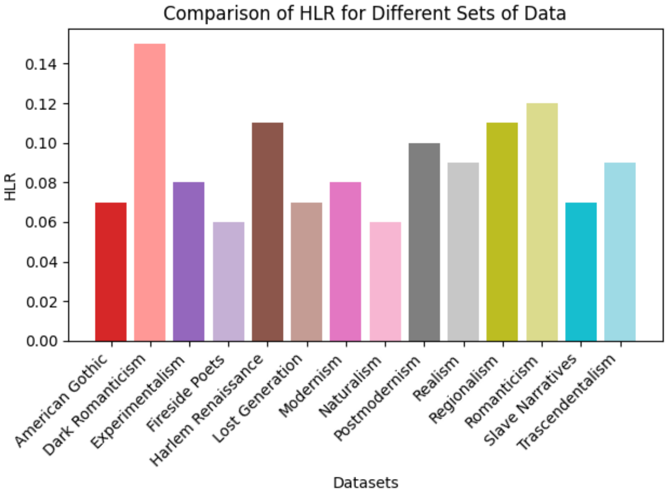

Figure 2: Hapax Legomena Ratio

6.3 Cosine Similarity between Movements

Extensive corpus analysis was also computed through cosine similarity among various literary movements based on TF-IDF vectorization. However, it's worth noting that this particular analysis carries less weight than the others due to the imbalance in the number of texts within each corpus, which impacts representational richness. Unlike other metrics that didn't involve direct similarity computations between corpora, this imbalance affects the overall results. Two standout findings include the high similarity with Transcendentalism and the low similarity with the Harlem Renaissance (Fig 4). Concerning the former, it's noteworthy how Gothic, in direct contrast to Transcendentalism from a philosophical perspective, shares a semantic space with common themes and subject terms, although the descriptions differ significantly. Conversely, the Harlem Renaissance (Fig 3), known for its political engagement and folk/journalistic style, deviates from the detailed descriptions found in Gothic literature. The themes explored range from social dynamics and subcultures seeking identity through folkloric anthropological analysis to introspective analyses delving into emotional and perceptual sensations related to the sublime essence of nature. Thus, this opposition reflects a fundamental contrast between introspective and socially outward-looking descriptions.

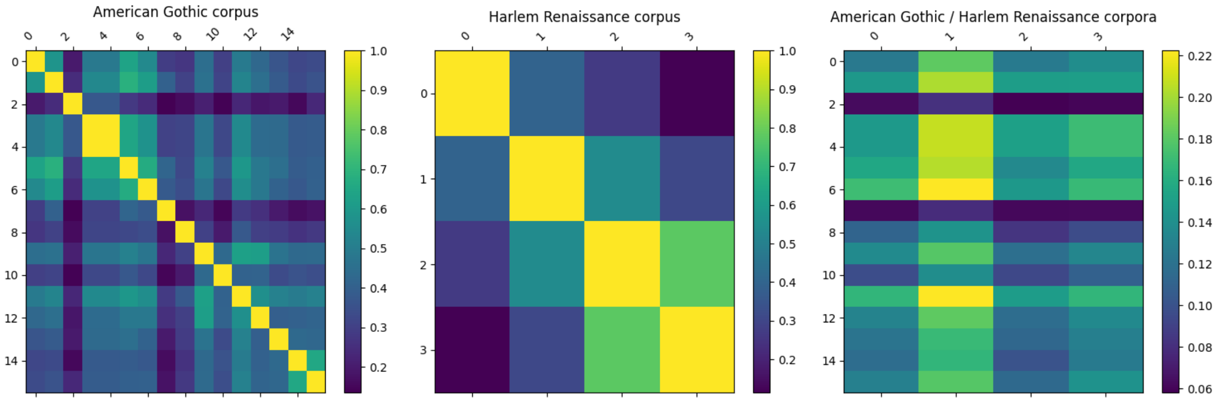

Figure 3: American Gothic:Harlem Renaissance

The final analysis focused on the internal similarity within the corpus, revealing interesting patterns (Fig 5). Notably, the Gothic, Dark Romanticism, and Romanticism movements exhibited low internal conformity, indicating a non-linear, original, and nonconforming style among their texts. It's worth emphasizing that Gothic, while still displaying low internal conformity, had slightly higher similarity compared to the other two, suggesting its evolution towards a more canonical style. In contrast, newer movements in American literature, such as the Harlem Renaissance and Slave

Narratives, showed greater internal variety. This underscores the diversity of themes explored by authors within these movements, reflecting their attention to different cultural and social conditions within American subcultures, resulting in distinct thematic elements in their works. The graphs of the results are in Appendix 2.

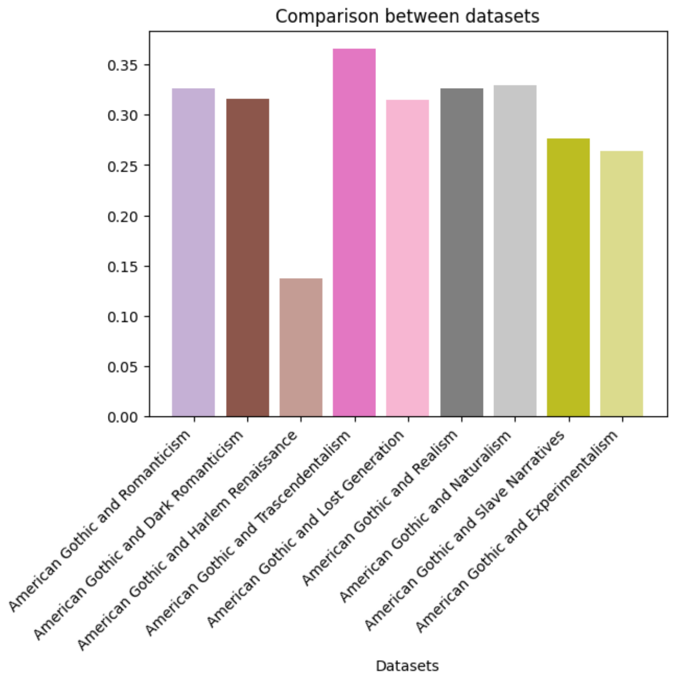

Figure 4: Comparison between Movements

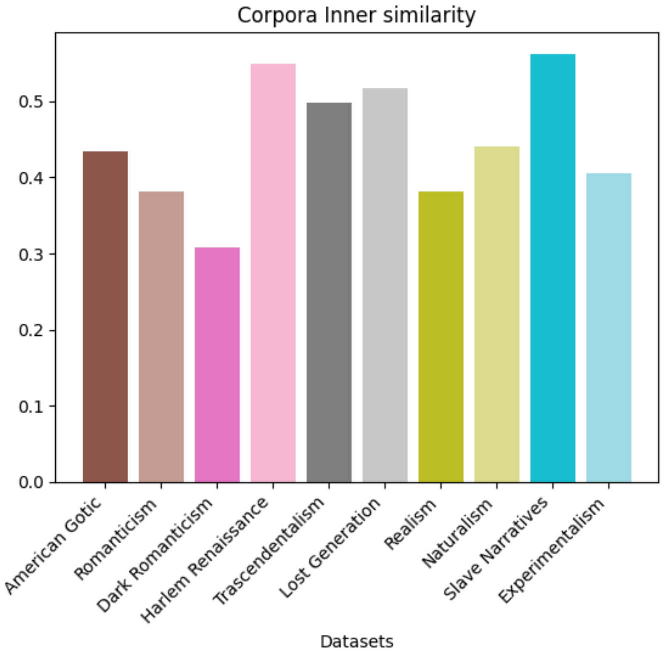

Figure 5: Movements Inner Congruity

7. Conclusion and Future Directions

My analysis, while open to more sophisticated methods, effectively demonstrates the fusion of distant and close reading techniques. By analyzing the Gothic movement and utilizing statistical patterns and frequency analysis, I have been able to extend classical hermeneutic critique attributes to a large number of texts and compare them with texts from various movements. The alignment between the attributes assigned to Gothic and the analysis results underscores the potential of this approach. It not only confirms existing insights but also offers a valuable complement to qualitative and interpretive analyses.

The project holds promising and various possible future explorations. Firstly, refining and expanding the representative canon, possibly by accessing platforms with a wider range of texts or enabling digital transmigration from print sources, would enhance the corpus consistency. Greater precision of canon is indeed essential for truly consistent attribution. Numerical normalization of the different corpora bodies would also allow more balanced analyses. The search for contextual correspondences could also provide another research opportunity, allowing the inclusion of not only content words but also grammatical words in the analysis. A delve into a syntactic domain analysis of this kind could reveal richer stylistic and narratological patterns. For instance, investigating prepositions and pronouns, particularly in the Gothic genre, could provide insights into characters' journeys and social dynamics, for instance. Additionally, incorporating an evolutionary and diachronic dimension to the genre analysis, considering publication years and geographical factors, could shed light on historical-contextual influences within the corpora. These directions hold great potential for enriching the project's understanding of the genre's development and characterization.

Appendix

Appendix 1 - Tables of terms frequency [American Gothic]

These are the overall results of the TF-IDF top frequency words retrieval, in a window of 100 terms. They are distinguished by their grammatical class and manually classified in semantic classes. The numbers correspond to their TF-IDF values.

Verbal class

Progressive motion actions come 3.6569 pass 1.6387 appear 1.6722 bring 1.8230 leave 1.8707 return 0.8444 continue 0.8943 follow 0.8587 return 0.8444 reach 0.8387 run 0.7890 proceed 0.7524 arise 0.7341 walk 0.5645 arrive 0.4937

Research find 4.2782 know 3.3191 discover 0.7883

Stasis stand 1.4017 remain 1.1241 sit 1.2159 write 1.1034 observe 1.0791 lead 0.6686 maintain 0.5472

Dialogic dynamics think 3.0099 tell 1.9041 appear 1.6722 speak 1.5712 turn 1.5280 meet 0.9872 ask 0.5553 mention 0.5518 admit 0.5067 suggest 0.5016 reply 0.4885 talk 0.4331 listen 0.4772

Adventurous turn 1.5280 fall 1.4818 throw 1.3719 hold 1.1770 open 1.0799 receive 1.0241 enter 0.9953 meet 0.9872 hang 0.9424 carry 0.9222 live 0.9132 die 0.7865 seek 0.7166 use 0.6875 form 0.6669 prove 0.6083 lose 0.6044 love 0.5243

Growth grow 1.2483 learn 0.5327 afford 0.4555

Reflective/Psychological States suppose 1.1470 consider 1.2113 observe 1.0791 believe 0.9423 draw 0.9361 continue 0.8943 read 0.8880 understand 0.8867 forget 0.7488 remember 0.6036 wish 0.5059 imagine 0.4663

Perceptive States hear 2.0170 feel 1.8483 appear 1.6722 see 1.2364 perceive 0.6024 suffer 0.5412

Nominal class

Sublime night 1.1453 mind 0.9408 soul 0.7417 spirit 0.6917 dream 0.5045

Natural Elements air 0.8345 water 0.7184 tree 0.7055 earth 0.5733 sea 0.5531 wind 0.5333 fire 0.4834 horse 0.4803 star 0.4747 sun 0.4291 island 0.4284

Temporal Markers time 2.9044 day 2.1550 night 1.1453 year 1.0265 life 0.9624 world 0.9448 fact 0.8453 moment 0.7206 point 0.7198 end 0.6991 hour 0.6971 morning 0.5313

Size Descriptors length 0.8682 manner 0.7428 kind 0.6582 circumstance 0.4453 sound 0.4809

Places place 1.1739 way 1.1339 country 1.0298 door 0.9148 house 0.9085 point 0.7198 end 0.6991 room 0.6521 death 0.6348 distance 0.4341 wall 0.4618 abbey 0.4810 chamber 0.4992 hall 0.4958 village 0.5002 window 0.5039 city 0.5498 land 0.5539

Body Parts eye 1.6506 hand 1.2173 head 1.2118 foot 0.8379 arm 0.5556 heart 1.1811

Discoursive voice 0.5521 question 0.5140 person 0.6087

Social Characters squire 0.8395 friend 0.8380 lady 0.7384 gentleman 0.6416 family 0.7719 law 0.4429 king 0.4511 power 0.5221

Meta-narrative Elements poem 0.7495 character 0.6453 illustration 0.6589 book 0.6407 story 0.5051

Adjectival class

Reflective/Psychological States suppose 1.1470 consider 1.2113 observe 1.0791 believe 0.9423 draw 0.9361 continue 0.8943 read 0.8880 understand 0.8867 forget 0.7488 remember 0.6036 wish 0.5059 imagine 0.4663

Perceptive States hear 2.0170 feel 1.8483 appear 1.6722 see 1.2364 perceive 0.6024 suffer 0.5412

Sublime old 4.7031 wild 0.9652 ancient 0.9348 deep 0.8151 dead 0.6273 dark 0.6143 strange 0.6086 vast 0.5812 particular 0.5213 different 0.5123 huge 0.4755 sad 0.4571 broad 0.4364 bad 0.4345 melancholy 0.4080

Positive meaning great 4.9728 good 3.3000 young 1.6287 happy 0.7555 fine 0.7505 dear 0.5917 rich 0.5804 sweet 0.5802 original 0.4818 bright 0.4815 peculiar 0.4790 favorite 0.4736 close 0.4703 light 0.4663 well 0.4660 distant 0.4611 sad 0.4571 ordinary 0.4383

Places descriptors little 3.9306 long 1.8632 high 1.3840 large 1.3518 small 1.3055 new 1.2588 open 1.0799 black 0.8738 beautiful 0.8542 low 0.8223 short 0.7425 white 0.6858 natural 0.6056 simple 0.5962 green 0.5681 similar 0.5599 single 0.5231 public 0.5139 fancy 0.5022 clear 0.4974 red 0.4873 half 0.4464 modern 0.4433

Adventurous worthy 0.6755 right 0.6742 late 0.6420 strong 0.6104 impossible 0.5670 necessary 0.5591 ill 0.5499 ready 0.5406 able 0.5046

Assertives of Presence present 1.3836 true 1.2822 certain 1.0024 sure 0.8362 possible 0.8349 proper 0.5469

Historical Features dutch 0.6659 english 0.4869 gentle 0.4375 noble 0.4225 french 0.3992

6.3.1 Adverbial class

Time and Movement long 4.1160 far 3.9961 soon 3.4985 away 3.1257 immediately 2.2600 suddenly 1.4548 gradually 1.0755 afterward 1.0645 finally 0.8472 actually 0.8311 continually 0.7041 slowly 0.6683 quietly 0.5785

Discourse Markers (Dialogic) merely 2.1596 nearly 2.0901 somewhat 1.3976 especially 1.3875 occasionally 1.2587 evidently 1.2099 generally 1.2084 surely 0.7818 pretty 0.5903

Measurements merely 2.1596 nearly 2.0901 scarcely 1.3902 evidently 1.2099 generally 1.2084 gradually 1.0755 entirely 1.0408 completely 0.9720 exceedingly 0.9642 precisely 0.9017 fully 0.8120 sufficiently 0.7813 partly 0.7572 perfectly 0.7419 frequently 0.7340 distinctly 0.6953 equally 0.6426 extremely 0.6365 deeply 0.6071 properly 0.6005

Tables of terms frequency [Trascendentalism]

Nominal class

Temporal time 1.4152 day 1.3947 night 0.8230 light 0.4628 hour 0.4148 morning 0.4039 death 0.3893 fall 0.3518 year 0.7226 end 0.3239 age 0.3192 summer 0.2779 spring 0.2657

Natural Elements tree 1.2567 apple 1.0909 life 0.9095 wood 0.9077 water 0.9060 river 0.8284 lake 0.7281 shore 0.6314 land 0.6294 place 0.6217 nature 0.6005 canoe 0.5952 earth 0.5653 fruit 0.5149 sun 0.4865 wind 0.4841 mountain 0.4789 sea 0.4423 rock 0.4223 forest 0.4199 air 0.3850 fire 0.3834 bird 0.3790 leave 0.3713 moose 0.3672 field 0.3642 island 0.3407 grass 0.3228 star 0.3178 fish 0.2813 hill 0.2491 sky 0.2443

Spirituality soul 0.5962 eye 0.5047 state 0.4930 stream 0.4924 sound 0.4026 spirit 0.2528

Body Parts foot 0.6789 hand 0.6612 head 0.4417 body 0.3994 face 0.3449 mind 0.2797

Historical/Social Features indian 0.4992 country 0.3952 captain 0.3952 ship 0.3872 boat 0.3836 war 0.3685 government 0.3283 law 0.3278 companion 0.2809

6.4 Appendix 2

Appendix 2 is inserted in the project repository on GitHub for practicality reasons. It contains two folders within the graphs of the cosine similarities computing and of the different complexity measures across the corpus. The link to the repository is the following: `https://github.com/memonji/Computational_Linguistics_2022-23.git`.

References

Abrams, M. H. and Harpham, G. (2014). A glossary of literary terms. Cengage learning.

Adolphs, S. (2006). Introducing electronic text analysis: A practical guide for language and literary studies. Routledge.

Alghamdi, R. and Alfalqi, K. (2015). A survey of topic modeling in text mining: LSA, LDA topic modeling, and topic evolution model. International Journal of Advanced Computer Science and Applications, 6(1), 21.

Balech, S.and Benavent, C. (2019). Les techniques du NLP pour la recherche en sciences de gestion. (hal-02400308).

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Boyles, N.and Scherer, M. (2012). Closing in on close reading. On Developing Readers: Readings from Educational Leadership, EL Essentials, 89-9.

Burrows, J. F. (1987). Computation into criticism: A study of Jane Austen's novels and an experiment in method.

Cannella, C. (2015). L'Estetica della Natura nel Trascendentalismo americano.

Conférence de Caroline Sporleder, Directrice du Centre for Digital Humanities de l'Université de Göttingen (Allemagne), Professeure Invitée au Mois de Mars 2018 Par Le Labex TransferS et Thierry Poibeau (Lattice), sur le Thème: Natural Language Processing for Digital Humanities.

Eve, M. P. (2022). The digital humanities and literary studies (p. 208). Oxford University Press.

Gómez-Adorno, H., Posadas-Duran, J. P., Ríos-Toledo, G., Sidorov, G., Sierra, G. (2018). Stylometry-based approach for detecting writing style changes in literary texts. Computación y Sistemas, 22(1), 47-53. doi: 10.13053/CyS-22-1-2882.

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Literary Movements. (n.d.). StudySmarter UK. Retrieved from https://www.studysmarter.co.uk/explanations/english-literature/literary-movements/.

Lvoff, B. (2021). Distant Reading in Russian Formalism and Russian Formalism in Distant Reading. Russian Literature, 122, 29-65.

Moretti, F. (2000). Conjectures on world literature. New left review, 2(1), 54-68.

Moretti, F. (2013). Distant reading. Verso Books.

Perazzini, F. (2013). Il Gotico@ distanza: Nuove prospettive nello studio dell'evoluzione dei generi del romanzo. Edizioni Nuova Cultura.

Pozzo, R. (2022). Recensione: Moretti, Franco. 2020. A una certa distanza: leggere i testi letterari nel nuovo millennio. Umanistica Digitale, (11), 231-236.

Project Gutenberg. (n.d.). Retrieved from http://www.gutenberg.org. Accessed 2023 January.

Rockwell, G., Sinclair, S. (2022). Hermeneutica: Computer-assisted interpretation in the humanities. Cambridge, MA: MIT Press.

Selden, R. (1995). From formalism to poststructuralism.

Sievert, C., Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces (pp. 63-70).

Sinclair, S. Rockwell G. (2013). The Art of Literary Text Analysis. GitHub. Accessed: [2023, February].

Sinclair, S. Rockwell G. (2018). The Art of Literary Text Analysis. Accessed: [2023, February].

Van den Rul, C. (2023, February). [NLP] Basics: Measuring The Linguistic Complexity of Text. Retrieved from Towards Data Science website: https://towardsdatascience.com/linguistic-complexity-measures-for-text-nlp-e4bf664bd660.

Voyant Tools https://voyant-tools.org/. Accessed: (2023) February.

Schiuma G. Carlucci D. (2018). Big data in the arts and humanities theory and practice. Auerbach Publications. https://doi.org/10.1201/b19744.