Quantifying the Information Flow of Long Narratives: A Case Study of Jane Austin's Works

Tianyi Zhang

School of International Studies, Zhejiang University, China

DOI : https://dx.doi.org/10.47772/IJRISS.2024.808067

Received: 22 July 2024; Accepted: 26 July 2024; Published: 31 August 2024

ABSTRACT

This study employs digital humanities to analyze the information flow in Jane Austen's classic literature using the GPT-2 XL model. Entropy, a measure of unpredictability, quantifies the narrative's dynamic engagement with readers. By calculating the entropy of each sentence, the research reveals unique patterns of information gain across Austen's novels, reflecting the ebb and flow of reader surprise. Peaks in entropy correspond to narrative climaxes, while declines indicate more predictable plot developments. The findings suggest that digital tools can offer fresh insights into literary analysis, highlighting the interplay between predictability and surprise in narrative structure. This exploratory approach to literature enriches traditional literary studies and opens new avenues for understanding reader engagement with classic texts.

Key words: information flow, entropy, narrative dynamics, Jane Austen, large language model

INTRODUCTION

Digital humanities, a cross-disciplinary field that integrate digital tools provided by computers and non-digital ones provided by traditional humanity researches, is revolutionizing researches in social humanities, and literature study is of no exception(Wilkens, 2015). In an information-explosion era, the mushrooming of various literature works fail to eradicate role and impact of classics in literature studies. However, traditional literature studies emphasize more on subjective interpretation in close-reading of individual works by quantitative means(Moretti, 2013), instead of observing the dynamism of information flow within and across texts in qualitative ways.

The largest amount of information transmitted through a communication channel could be measured by entropy,which represents the average information content and the average uncertainty of a discrete variable (Cowan, 2000; Lambert, 2004). Assume a text is a random variable T consisting of tokens from a set of units W = {W1, W2, …, Ww}, the word entropy of T can be calculated as (Shannon, 1948):

$$H(T) = -\sum_{i=1}^{w_i} \quad p(wi)\ \log2\ p(wi) \qquad (1)$$

The higher entropy is associated with more uncertainty about the value of W, suchthat entropy is maximal when all possible values of W have the same probability, andentropy is zero when there is 100% certainty about the value.

We perceive information differently when reading. Information across narratives is distributed unevenly in local and global levels, so that we can be grabbed by the content with weighing up unpredictability against predictability. In other words, the arrangement of plots in narration presents dynamic amount of information to attract the reader by making contents less or more predictable (Kukkonen, 2014).

In reading, our brain automatically predicates various inputs at the moment of information receiving. We are less surprised with successful predication and are surprised when the content is out of expectation. The degree of predictability could be calculated by Large Language Models (LLMs) in terms of entropy. LLMs are a new family of language models (LMs) represented as large-scale neural networks, which have rapidly come to serve as the foundation of most current natural language processing systems (Bommasani et al., 2022). This exploratory study utilizes transformer-based GPT-2 Model (Radford et al., 2019) to calculate the entropy of each sentence of a long narrative, and all words in each sentence will be analyzed to visualize the information flow.

Good narratives have certain manners of information flow to grab our attention, so that we are engaged in the story during different levels of surprise. Classical literary works of Jane Austin are good cases to investigate the information flow of long narrative texts. The hypothesis is that each part of the long narratives would have distinctive information flow. The use of GPT-2 Model to investigate literature studies conforms to the major trend of developing digital humanities. In doing so, we innovate traditional literary researches and refresh our understanding of literature works and cognitive processes.

METHOD

A. Entropic Measures of Structure

Entropy quantifies the degree of uncertainty about what is being communicated as a sentence unfolds. Entropy fluctuates as we encounter each new word, with incomingwords affecting expectations regarding what will come next (Lowder et al., 2018). The amount of gained information at each input can be measured by entropy reduction(Hale, 2006, 2016). If entropy is reduced from one word to the next, then communicative uncertainty hasbeen reduced. Entropy reduction capture unique aspects of information complexity and, as such, should both serve as useful metrics for quantifying word-by-wordpredictability during incremental sentence processing(Hale, 2016). Information gain is also referred to as relative entropy for the amount of change in a probability distribution (Kullback&Leibler, 1951). For a set of units W = {W1, W2, …, Ww}, information gain is calculated as:

$$\mathrm{D_{kL}}(\mathrm{P}| \quad |\mathrm{Q}) = \sum\nolimits_{\mathrm{w}\in\mathrm{W}} \mathrm{P}(\mathrm{w}) \log \quad \left(\frac{\mathrm{P}(\mathrm{w})}{\mathrm{Q}(\mathrm{w})}\right) \ (2)$$

where P is the updated probability distribution when encountering one word $\mathrm{P}(\mathrm{w_t}|\mathrm{w}_{1\ldots\mathrm{t}\cdot})$ in the set and Q is the original distribution $\mathrm{P}(\mathrm{w_t}|\mathrm{w}_{1\ldots\mathrm{t}-1})$.

The predictability of linguistic units in texts or sentences can be directly observed in self-paced reading times by eye-tracking. There is growing evidence that information gain or entropy reduction is in fact a significant predictor of sentence-processing times. Several studies have now shown that higher entropy reduction contributed to increased reading times (Wu et al., 2010; Frank, 2013; Linzen& Jaeger, 2016; Lowder et al., 2018). In such sense, information gain or entropy reduction can reliably mirror the predictability of words during sentence processing.

B. The Use of LLM to Quantify Information Flow

GPT-2 XL are used in this study. It is a 1558 million-parameter neural LLM trained on over 6GB of text data (Sap et al., 2022). Through self-supervision learning, it grasps the likelihoods of all linguistic units in certain context. By assigning probabilities of every single word of texts with consideration of all preceding words within a calculation window, it can simulate brain's function to make predications during reading.

The composition of narrative texts involves information flow, where the construction of one message follows certain logic with references are made to previous parts as context needed to understand later parts of the same message(Estevez-Rams et al., 2019). All sentences in a story are dedicated to a certain topic. Context is imperative in understanding story plot, and the presence of one sentence serves as information gain to influence our predication of the following. Such can be referred to as the sequentiality of story: each sentence is generated conditioned on the story topic as well as all of its preceding sentences(Sap et al., 2022). Sequentiality leverages probabilities of words and sentences in stories to determine the difference in the likelihood of sentences given the preceding information. It provides a measure of narrative flow based on probabilities of story sentences, which can be estimated by large language models (LLMs).

On a micro level, the amount of information of the whole sentence is calculated by adding all the information gain of each words considering the presence of all the preceding words. In one sentence, the presence of the preceding word determines the presence of the next. In other words, the occurrence of previous ones reduces the uncertainty of the next, so that each word in sentence is assigned with unique probability of occurrence due to the previous information. The amount of uncertainty reduction equals information gain, which can be calculated by the GPT-2 model (the first word is an exception, and its probability is calculated based on the distribution of the whole training data). On a macro level, the information gain or sequentiality of one sentence takes the previous context, or all preceding sentences, in to account. One long narrative of about 100,000 words can be composed of roughly 7,000 sentences. Considering the calculation price, this study only takes one sentence as the preceding context. Sequentiality is measured for each sentence Si of the long narrativesand for the targeted sentence and its preceding sentence as the difference in the negative log-likelihood (NLL) of the sentence:

$$I_{(s_i)} = -\log_P(s_i) + p\log_P(s_i|s_{i-1})(3)$$

First, single sentence as separated by quotation marks is taken as a window to be calculated separately. Then, the sentence and its previous sentence are taken into the window in a gliding way. If a text contains 7,000 sentences, the first step generated 7,000 values, and the second step generated 6,999 values. Value of the latter minus value of the former correspondingly equals the information gain of each sentence. Furthermore, since the first sentence has no preceding context, it serves as base value. Information flow of the whole narratives is then displayed by calculating the corresponding information gain of each sentence.

C. The shape of narratives

Stories have a generalized narrative arc in the building of the story's scene(Freytag & MacEwan, 1960). As a story moves forward, action between characters increases and, ultimately, peaks at the top of the narrative arc: the story's climax. Subsequently, a decline in conflict prompts characters to transition toward the denouement or the resolution. The progression of plot is a dynamic process. Narrative stories have unique shapes in progressing plots. The arrangement of plot has its own pace and timing. It coordinates characters' movement in the story world with the progress of the narrative itself. (Laurino Dos Santos & Berger, 2022). Story tellers utilize narrative elementsto maintain the audience's engagement and memory of stories, which lead to an information progression dynamism. To quantify such progression, long narratives of Jane Austin are broken into 5 approximately equal-sized chunks as the progression of whole stories. Information flow within each part is computed to visualize the progression and thus reveal the shape of stories.

D. Materials

All six classics of Jane Austin are selected from the Gutenberg Project: Pride and Prejudice, Persuasion, Sense and Sensibility, Emma, Northanger Abbey and Mansfield Park. Word counts of nine bildungsroman range from 101,678 to 210,254 and thus belong to long narratives.

E. Procedures

The first step is material preprocessing. All texts are pre-processed by NLTK tools. After removing special characters and empty lines, texts are tokenized and are segmented into sentences based on common punctuation marks "!", ".", "?"and ";".

The second step is calculation. Each sentence is inputted into GPT-2 XL model provided by Hugging face. The model calculates the average entropy of a given sentence by assigning probability of each word in sentences given its preceding words.

The last step is data analysis. Information gain of each sentence is recorded, as computed by LLM in terms of entropy. Since different sentences have different length, the average information gain per token in one sentence is computed and recorded. All entropy values are grouped into 5 blocks in a liner way to indicate plot progression, with each block contains roughly same number of sentences to visualize the information flow. Average sentence length as measured by token sizes of sentences and the average entropy per token of all sentences are used as indicators of cross-text differences.

All selected materials are divided into sentences and inputted into the model, which then generate the entropy values of sentences. The whole processes are conducted in python. Data of materials are collected and analyzed to display the information flow.

RESULTS

All six long narratives of Jane Austin were analyzed. Stories were of different lengths, with Emma being the longest one and was segmented into 9,528 sentences, and Northanger Abbey being the shortest one and was segmented into 4,976 sentences. Though the stories were of various lengths, the average sentence length and average entropy per token of these works were similar, ranging from 21.43-22.55 and 0.44-0.54 respectively. Sense and Sensibility had the longest sentence length in, while Mansfield Park is the shortest at 21.43 tokens per sentence.

The six novels were of starkly different lengths. To facilitate analysis, all the information gain of sentences were grouped linearly into blocks indicate plot progression by 20%, 40%, 60%, 80% and 100%, with each block contains roughly same number of sentences to visualize the information flow. As manifested in Figure 1, the result visualized the shape of different novels in unique patterns. Sense and Sensibility, Persuasion, Northanger Abbey and Mansfield Park displayed a rise-fall information flow in general. However, Pride and Prejudice presented a different story: the information gain decreased when the plot was progressed by 40%, and then when upward. Information flow in Emma was similar to Pride and Prejudice in the beginning. However, after the story marched toward 80%, the overall gain decreased. Besides, most of the peaks occurred in the middle of the plot progression, and the beginning and ending were more likely to have lower information gain.

DISCUSSION

Narratives unfolds across time as narration extends itself through time. Readers thus get involved in particular stretches of events in the plots by word following word and sentence following sentence. When reading, information carried by plots in the form of sentence are perceived differently by readers. Plot progression is a dynamic process, so that readers feel the attractiveness of the story when they receive dynamic amount of information.

On the one hand, dynamic information flow pattern proves the attractiveness of stories. The result of information flow displays unique shapes of stories with information peaks and lows during plot progression. Jane Austin's works are renowned for talented plot arrangements, which vividly depict characters and deliver thought-provoking ideas. The longstanding attractiveness of her plot progression could be mirrored in an interplay of ups and downs in information gain, where the predictability and unpredictability deliver dynamic amount of information to attract the reader(Kukkonen, 2014).

On the other hand, the result indicates commonalities and differences of story shape. As for the common ground, information flow pattern is dynamic in good narratives like Jane Austin's work. Inconsistency of flows across progression means that we are engaged in the story during different levels of surprise. In reading, our brain automatically predicates various inputs at the moment of information receiving. We are less surprised with successful predication and are surprised when the content is out of expectation. Readers can feel carried along by the different information gain in sentences in the narrative. As shown in the information flow patterns, an increase means more surprise, and a decrease in turn equals less surprise. The amount of information is presented in various manner, so that readers are constantly activated by the content.

As for the differences, novels may present information in a different manner. Though stories have a generalized narrative arc in the building of the story's scene(Freytag & MacEwan, 1960) in terms of beginning, progressing, climax and ending, such format does not mean a similar distribution pattern of information.

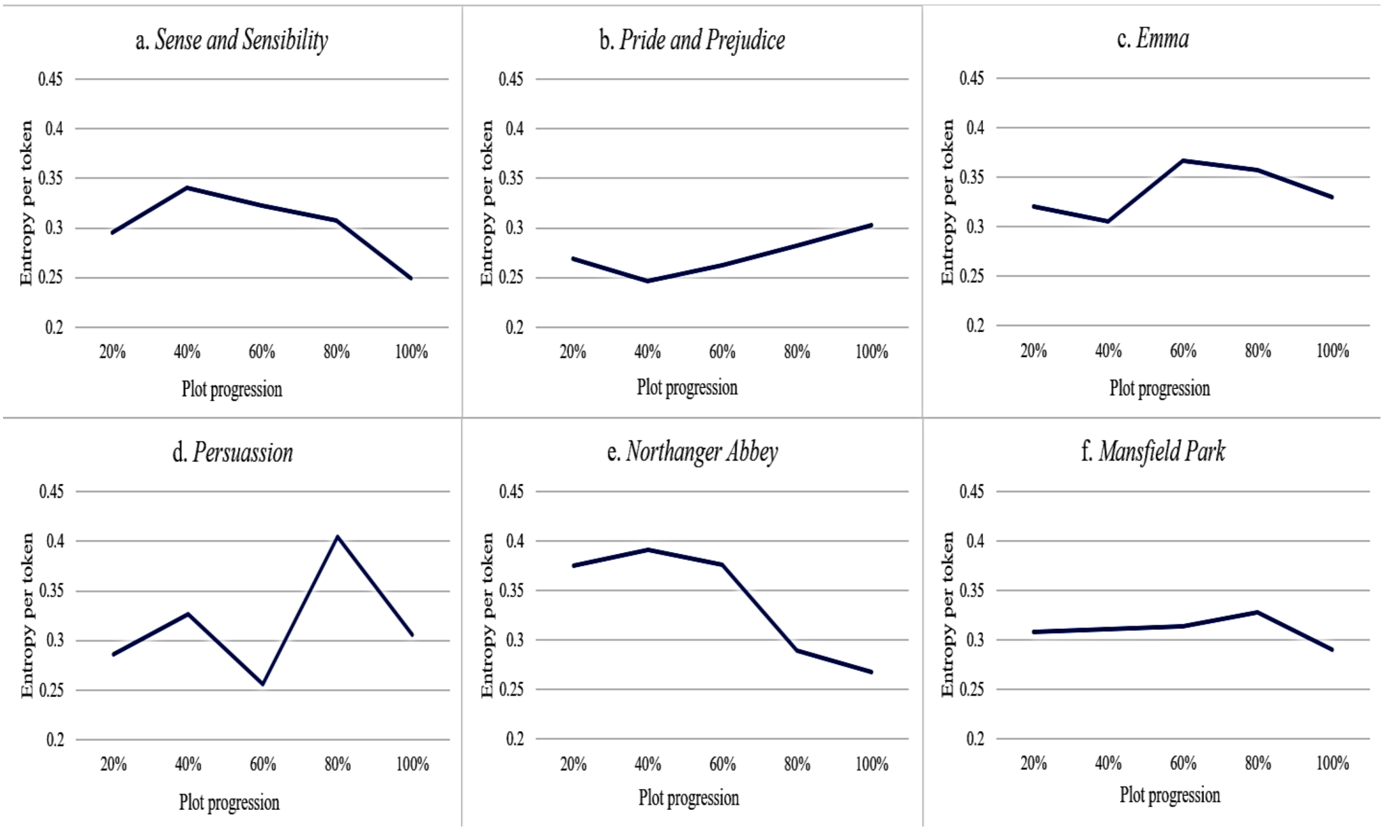

Fig.1: Information Flow of Jane Austin's Work

Table 1: Jane Austin's Long Narratives

Title	Token Count	Sentence Count	Average Sentence Length	Average Entropy per Token
Sense and Sensibility	146,427.00	6,493	22.55	0.3
Pride and Prejudice	161,177	7,513	21.45	0.27
Mansfield Park	193,356	9,099	21.25	0.31
Emma	210,254	9,528	22.07	0.34
Persuasion	109,613	4,976	22.03	0.32
Northanger Abbey	101,678	4,744	21.43	0.34

As shown in the flow pattern, Sense and Sensibility, Persuasion, Northanger Abbey and Mansfield Park displayed a rise-fall information flow in general. During the plot progression, readers can feel more surprised at first, and feel less surprised in the end. However, when it comes to Pride and Prejudice, readers may be more capable of predicting the content when the plot was progressed by 40%, and then find the plot become more attention-grabbing afterwards since the information gain goes down. Information flow in Emma was similar to Pride and Prejudice in the beginning. Besides, a grasp at the information flow pattern can tell the most predictable and the most unpredictable parts in plot progression. The beginning and the ending plots are more predictable given the relatively lower information gain values, and the middle part, namely the 40% to 80% part display more uncertainty given the relatively higher values. Such patterns also correspond to our common knowledge that climax in the stories contains more conflicts and is thus more unpredictable. Furthermore, readers may have the same general feeling of being able to predict the plot progression when read Sense and Sensibility and Mansfield Park since the peak gain happens in the 20% and the valley gain takes place at the final part.

It worths noticing that all values were computed based on the internal logic of large language models. On the one hand, while sentences were separated by common punctuation marks including "!", ".", "?"and ";", the model took the presence of all marks into consideration when calculating the token size. When it computed entropy based on probability, the model presupposed that the occurrence of these marks also influenced the predictability of following content. On the other hand, the average entropy per token calculated by GPT-2 models was starkly different from the common results as calculated by QUITA, the result of which was about seven bit per token for Jane Austin's work(Shi & Lei, 2020). GPT-2 serial model was firstly trained based on big data, and it then decoded all presented materials into vectors multiple layers and assigned attention weights to each token based on their position through a masked self-attention block(Dai et al., 2019). In other words, it acquired probability of each token by projecting all tokens into a space which was much larger than the token sizes of single literature work. However, the use of computational linguistic software like QUITA only considered the text itself, which was not capable of simulating the function of human brain. To be specific, such huge difference was caused by different calculating mechanism.

CONCLUSION

This exploratory study utilizes GPT-2 Model to measure the information flow pattern of long narratives represented by Jane Austin's works. Information gain in terms of entropy values of all sentences of each narrative are grouped linearly in equal blocks to visualize the flow pattern. Each part of the long narratives has distinctive information flow patterns, which represent the predictability and unpredictability of contents. This piloted study is intended to refresh our understanding of literature works and cognitive processes under the digital Humanities by means of large language modeling. Further improvement is needed to combine the detailed content analysis with information flow pattern, and more literature works of other representative authors shall be included to provide more solid evidences of information flow pattern.

REFERENCES

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258
2. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context (arXiv:1901.02860). arXiv. https://doi.org/10.48550/arXiv.1901.02860
3. Estevez-Rams, E., Mesa-Rodriguez, A., & Estevez-Moya, D. (2019). Complexity-entropy analysis at different levels of organisation in written language. PLOS ONE, 14(5), e0214863. https://doi.org/10.1371/journal.pone.0214863
4. Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. Topics in Cognitive Science, 5(3), 475–494. https://doi.org/10.1111/tops.12025
5. Freytag, G., & MacEwan, E. J. (1960). Technique of the Drama: An Exposition of Dramatic Composition and Art. https://www.semanticscholar.org/paper/Freytag's-Technique-of-the-Drama%3A-An-Exposition-of-Freytag/2882acb56b7917c2ddfe9b31d08cbf7f6fdb9031
6. Hale, J. (2006). Uncertainty about the rest of the sentence. Cognitive Science, 30(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64
7. Hale, J. (2016). Information ‐ theoretical Complexity Metrics. Language and Linguistics Compass, 10(9), 397 – 412. https://doi.org/10.1111/lnc3.12196
8. Kukkonen, K. (2014). Bayesian Narrative: Probability, Plot and the Shape of the Fictional World. Anglia, 132(4). https://doi.org/10.1515/ang-2014-0075
9. Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694
10. Lambert, S. (2004). Shared Attention during Sight Translation, Sight Interpretation and Simultaneous Interpretation. Meta : Journal Des Traducteurs / Meta: Translators’ Journal, 49(2), 294–306. https://doi.org/10.7202/009352ar
11. Laurino Dos Santos, H., & Berger, J. (2022). The speed of stories: Semantic progression and narrative success. Journal of Experimental Psychology. General, 151(8), 1833–1842. https://doi.org/10.1037/xge0001171
12. Linzen, T., & Jaeger, T. F. (2016). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. Cognitive Science, 40(6), 1382–1411. https://doi.org/10.1111/cogs.12274
13. Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. Cognitive Science, 42(S4), 1166–1183. https://doi.org/10.1111/cogs.12597
14. Moretti, F. (2013). Distant Reading. Verso Books.

15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe
16. Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. Proceedings of the National Academy of Sciences, 119(45), e2211715119. https://doi.org/10.1073/pnas.2211715119
17. Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
18. Shi, Y., & Lei, L. (2020). Lexical Richness and Text Length: An Entropy-based Perspective. Journal of Quantitative Linguistics, 29, 1–18. https://doi.org/10.1080/09296174.2020.1766346
19. Wilkens, M. (2015). Digital Humanities and Its Application in the Study of Literature and Culture. Comparative Literature, 67, 11–20. https://doi.org/10.1215/00104124-2861911
20. Wu, S. T., Bachrach, A., Cardenas, C., & Schuler, W. (2010, July 11). Complexity Metrics in an Incremental Right-Corner Parser. Annual Meeting of the Association for Computational Linguistics. https://www.semanticscholar.org/paper/Complexity-Metrics-in-an-Incremental-Right-Corner-Wu-Bachrach/5cfe9bb78fa50955be5a99e6966dd5ba5b4d4f80