Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Perplexity Games: Maoism vs. Literature through the Lens of Cognitive Stylometry

Maciej Kurzynski[1]
maciej.kurzynski@ln.edu.hk

[1]The Advanced Institute for Global Chinese Studies, Lingnan University, Hong Kong

Abstract
The arrival of large language models (LLMs) has provoked an urgent search for stylistic markers that could differentiate machine text from human text, but while the human-like appearance of machine text has captivated public attention, the reverse phenomenon—human text becoming machine-like—has raised much less concern. This conceptual lag is surprising given the ample historical evidence of state-backed attempts to regulate human thought. The present article proposes a new comparative framework, Perplexity Games, to leverage the predictive power of LLMs and compare the statistical properties of Maospeak, a language style that emerged during the Mao Zedong's era in China (1949-1976), with the style of canonical modern Chinese writers, such as Eileen Chang (1920-1995) and Mo Yan (1955-). The low perplexity of Maospeak, as computed across different GPT models, suggests that the impact of ideologies on language can be compared to likelihood-maximization text-generation techniques which reduce the scope of valid sequence continuations. These findings have cognitive implications: whereas engineered languages such as Maospeak hijack the predictive mechanisms of human cognition by narrowing the space of linguistic possibilities, literature resists such cognitive constraints by dispersing the probability mass over multiple, equally valid paths. Exposure to diverse language data counters the influences of ideologies on our linguistically mediated perceptions of the world and increases the perplexity of our imaginations.

keywords
perplexity, linguistic engineering, ideological discourse analysis, China, predictive processing

I INTRODUCTION

The problem of predictability in language is meaningful for humans and machines alike, the heated debates over the differences between natural and artificial intelligence notwithstanding (Piantadosi [2023]; Kodner et al. [2023]; Chomsky [2023]). Already in 1951, Claude Shannon's study "Prediction and Entropy of Printed English" blurred the distinction between the two by posing an extraordinary question: "How well can the next letter of a text be predicted when the preceding *N* letters are known?" (Shannon [1951, 50]). Shannon argued that human speakers possess a vast, implicit knowledge of language statistics which enables them to "fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation" (Shannon [1951, 54]). This computational perspective has been crucial for the advancements in natural language processing over the last few decades, but the implications of his observation that not all writing forms are equally predictable have drawn significantly less attention. Traditionally, stylometrists have focused on function words, syntactic structures, and thematic elements (Lutosławski [1898]; Mosteller and Wallace [1963]; Burrows [1987]; Herrmann et al. [2015]; Rybicki et al. [2016]; Evert et al. [2017]), whereas "predictability" has been rarely

considered a valid stylistic feature (Nikolaev et al. [2020]; Bizzoni et al. [2020]).[1] Only recently has the ability of large language models (LLMs) to generate human-like output spurred an urgent search for stylistic markers that could differentiate machine text from human text, with word predictability remaining one of the few viable candidates (Wu et al. [2023]; Chakraborty et al. [2023]; Mitchell et al. [2023]).

While the human-like appearance of machine text has captivated public and academic attention, however, the reverse phenomenon—human text becoming machine-like—seems to have raised much less concern. One reason for this conceptual lag might be the persistent assumption that human creativity inherently resists quantification, an assumption which can be easily qualified given the abundant historical evidence of state-mandated attempts to regulate human thought. The foundational studies in the field of ideological discourse analysis (IDA), such as works by Cameron [1995], Klemperer [2006], and Epstein [1991], suggested the potential of statistical approaches to the human/machine problem but did not fully utilize them. A notable exception and a major inspiration for the present article is Ji Fengyuan's study *Linguistic Engineering: Language and Politics in Mao's China* (Ji [2003]), which demonstrates how the Chinese Communist Party, under Mao Zedong's leadership, employed linguistic mechanisms to transform the newly established People's Republic of China (PRC). This involved introducing new political vocabularies, redefining traditional terms, and suppressing words contradictory to Party objectives. Importantly, Ji does not focus only on the morphological aspects of the ultra-politicized discourse; she also argues that tinkering with the modern Chinese language induced multiple psychological effects, such as associative priming and higher-order conditioning, which made select political terms and semantic associations more readily available than others in the revolutionary brain.

Linguistic Engineering was published when the computational resources required by LLMs were not yet publicly available. Since then, the growing availability of large textual corpora and the advent of Transformer models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) have revolutionized the study of human language (Vaswani et al. [2017]; Radford et al. [2018]; Devlin et al. [2018]). Within the context of computational stylometry, LLMs have enhanced authorship attribution (Fabien et al. [2020]; Rivera-Soto et al. [2021]), allowed scholars to build style embeddings and authorship models (Patel et al. [2023]; Wang, A. et al. [2023]; Huang, W. et al. [2024]), and even learned behavioral styles from chess games (McIlroy-Young et al. [2022]). As such, LLMs have the potential to replace traditional stylometric methods focused on word frequencies, syntactic patterns, or semantic coherence (Stamatatos [2009]; Neal et al., [2017]; Seroussi et al., [2014]; Sari et al., [2018]), as well as those based on older network architectures (Ruder et al. [2016]; Ou et al. [2023]). They can also shed new light on the human/machine dynamics, given the increasingly well-documented structural similarities between the multi-layered architectures of artificial neural networks and the human cognitive setup (Caucheteux et al. [2022, 2023], Chang [2022]). This article leverages such new computational resources to revisit Ji Fengyuan's insights, examining how ideologies increase predictability in texts and how texts resist ideologies by becoming unpredictable.

[1] According to most definitions, literary style does not impact the amount of transmitted information, which is inversely correlated with its predictability, but only *how* this information is transmitted. Among the six definitions of style provided by Herrmann et al. [2015] in their extensive literature review, there are only two which are broad enough to include predictability: (4) "an artifact that presupposes (hypothetical or factual) selection/choice among a set of (more or less synonymous) alternatives" and (6) "any property of a text that can be measured computationally." While the definition of style as "deviation from the norm" (5) touches on predictability, with the norm being arguably more predictable than the deviation, it has been usually adopted with respect to more directly observable features like word choice and sentence structure, rather than strictly statistical measures such as entropy and perplexity.

II DATASET

The dataset used in this study consists of four corpora: the selected works of Mao Zedong 毛泽东 (1893-1976), the collected modern prose[2] of Mo Yan 莫言 (1955-) and Eileen Chang 张爱玲 (1920-1995), and a larger compilation of 102 Chinese novels published by 62 writers active in the post-Mao era (Table 1). I treat Mao Zedong's writings as a proxy for Maospeak (*mao yu* 毛语, in its written form also called *mao wenti* 毛文体; Li [1998]; Link [2013]; Barmé [2012]), as it was chiefly through quotations from Mao that the discourse of class struggle and popular militarization spread across the PRC, thus shaping the everyday language.[3] This influence was particularly evident during the Cultural Revolution (1966-1976), when inability to quote the *Little Red Book*, the famous compilation of Mao Zedong's statements, could be taken as proof of reactionary (anti-communist) stance and lead to grave consequences (Ji [2003, 151], Schoenhals [2007]). I chose Mo Yan, the 2012 Nobel Prize laureate in literature, as he has been frequently accused by literary critics of inheriting the Maoist style in his lavish descriptions of war-time brutality (Link [2012]; Sun [2012]; Laughlin [2012]). Eileen Chang serves as another control writer: she spent most of her life outside mainland China after leaving the PRC in 1952, unable and unwilling to adapt to the communist regime. Critics noted Chang's unyielding efforts to avoid "any political dominion over literary creation by either side of the Cold War antagonism" (Wang, X. [2013, 295]), and her exposure to a wide range of conflicting worldviews in China, Hong Kong, Japan, and the United States led to a multifarious literary output deeply shaped by individual experience. Finally, the mixture of contemporary Chinese writing serves as yet another control corpus, offering a sample of modern literary Chinese. Mao's writings have been preprocessed by removing footnotes and lines of text shorter than 50 characters to filter out titles, dates, and signatures attached by editors to his letters and communiques. Other corpora have been cleaned of editorial comments, critical introductions, and footnotes. Since Chinese does not use spaces between words, all texts have been segmented (when needed) with the *spaCy* parser for Chinese.

Corpus	Tokens (Words)	Vocabulary (Unique Words)	Characters	Type-Token Ratio
Maospeak	1,685,940	55,469	2,894,023	0.0329
Contemporary	17,414,068	498,874	27,280,419	0.0286
Mo Yan	2,548,582	131,479	3,983,022	0.0515
Eileen Chang	822,809	58,770	1,254,864	0.0714

Table 1. Dataset statistics.

III EXPERIMENTS

3.1 Perplexity

The bulk of language model training involves iterating over a large amount of text and learning either to predict the next token given a sequence of tokens (GPT) or to reconstruct the original sequence given a perturbed one (BERT). These learned predictions can be then used to calculate the "surprisingness" of the actual words encountered in a previously unseen test dataset.

[2] "Modern prose" is an important distinction here, as Eileen Chang worked extensively on pre-modern Chinese literature such as the *Dream of the Red Chamber* 红楼梦, quoting the original text profusely in her critical essay *The Nightmare in the Red Chamber* 红楼梦魇 and related texts. Similarly, Mao Zedong composed classical-style poems. Such texts have been removed from the dataset, as my primary interest is in modern Chinese.

[3] While some of Mao's writings were not penned by him personally, the texts stamped with Mao's name remained most authoritative. Quoting Mao became an eristic (and at times, survival) strategy accompanying various language registers.

More technically, the training objective is to minimize the cross-entropy between the model's estimated probability distribution, Q, and the true probability distribution P underlying the training data. Since P is typically unknown, training leverages the empirical distribution, often represented by one-hot encodings for actual tokens, to approximate P.[4] This process effectively reduces the Kullback-Leibler (KL) divergence between Q and this empirical distribution, aligning the model's predictions more closely with the actual distribution of the language. One of the key intrinsic metrics for evaluating the learning process is perplexity (Huang et al. [2001]; Józefowicz et al. [2016]), calculated as the exponentiated average negative log-likelihood of the observed tokens in a given test sequence (1). For generative language models like GPT, the perplexity is derived sequentially by computing the probabilities for the next token given only the preceding (known) ones. The lower the perplexity, the more precise and confident the model's prediction capability.

$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log\ Q(y_i)\right) \quad (1)$$

Despite the well-documented negative correlation between perplexity and the model's familiarity with the training data, exactly what humanly interpretable feature of a text is conveyed by perplexity remains unclear. This black-box nature of language models did not prevent scholars from deriving new insights into human language. For example, perplexity has been used to distinguish transcripts of speech produced by patients with Alzheimer's disease from those produced by healthy individuals (Cohen and Pakhomov [2020]) and to measure diachronic language change (Pichel Campos et al. [2018]). Perhaps most pertinent to the present discussion are two papers by Miaschi et al. [2020, 2021], which show no inverse correlation between perplexity and readability, a metric which considers factors such as word difficulty and sentence length. This is a counterintuitive finding: a relatively simple text with high readability score ("The theft has taken place Thursday night"), variations of which should have been seen multiple times by a model during training, might be *more* perplexing to this model than a non-readable, highly technical text characterized by less variability ("the bioethics committee: no to euthanasia").[5]

3.1.1 Perplexity and Discursive Resistance

The following stylometric experiment follows up on Miaschi's insights by analyzing the impact of linguistic variability on the perplexity of language models. While the final perplexity score is undoubtedly a result of many contributing factors, here I focus on discursive predictability understood as a function of the number of possible word combinations (branching factor) in a sequence of tokens. Specifically, I used an open-source Chinese GPT-2 model with 30 million parameters and conducted a fine-tuning process at four distinct levels of variability:

- **Zero Variability**: The model was fine-tuned using only the original Chinese sentence 我们喜欢这个地方 ("We like this place"), without any alterations. This scenario served as a baseline to understand the model's performance while training on static input.
- **Low Variability**: Minor variations were introduced to training data by randomly sampling the grammatical object. Using the template 我们喜欢这个 [object] ("We like

[4] A one-hot encoding is a group of bits in which there is only one high (1) bit, all the others being low (0). In language modeling, this means that the actual word will be encoded as 1, with all other words in the vocabulary marked as 0.
[5] Both examples (translations from Italian) come from Miaschi et al. [2020].

this [object]"), I generated new training sentences, such as 我们喜欢这个公园 ("We like this park") or 我们喜欢这个图书馆 ("We like this library").

- **High Variability**: More variation was introduced by randomly sampling subjects, verbs, and objects using the template "[subject][verb] 这个 [object]." This resulted in a diverse set of training sentences, such as 我讨厌这个博物馆 ("I dislike this museum") or 他们建设这个学校 ("They build this school").
- **Mixed Variability**: During the first 20% of the training, the model was presented with unchanged sentences (as in the zero-variability scenario). Then, low variability was introduced for the following 40% and high variability for the last 40%.[6]

Before the start of each scenario, the model was reloaded and brought to its original pre-trained state. During fine-tuning, the model would see 1,000 training sentences, split into batches of 10 to stabilize the gradient descent. In each batch, the model would first attempt to predict the training sentences (with zero, low, or high variability, depending on the scenario). The loss would be then calculated and the weights updated by the AdamW optimizer. After each update, the model's perplexity on the original, unperturbed sentence 我们喜欢这个地方 ("We like this place") would be computed and recorded. This approach provided a controlled mechanism to simulate different degrees of linguistic variety and monitor the resulting shifts in perplexity.

Corpus	Substitutions
Subjects	我, 你, 她, 他, 我们, 你们, 他们, 她们, 咱们
Objects	喜欢, 讨厌, 考虑, 明白, 思考, 分析, 塑造, 创作, 记住, 参观, 离开, 返回, 发现, 爱上, 建设, 破坏, 使用, 访问, 忘记, 改变, 开始, 结束, 增加, 减少
Verbs	地方, 餐厅, 公园, 图书馆, 电影院, 工厂, 学校, 书店, 旅馆, 实验室, 大学, 厨房, 体育馆, 展览, 大排档, 教堂, 寺庙, 海滩, 博物馆, 音乐会, 咖啡馆, 市场, 花园, 河流, 超市, 宠物店, 游乐场, 水族馆, 画廊, 剧院

Table 2. Words used for sentence perturbations.

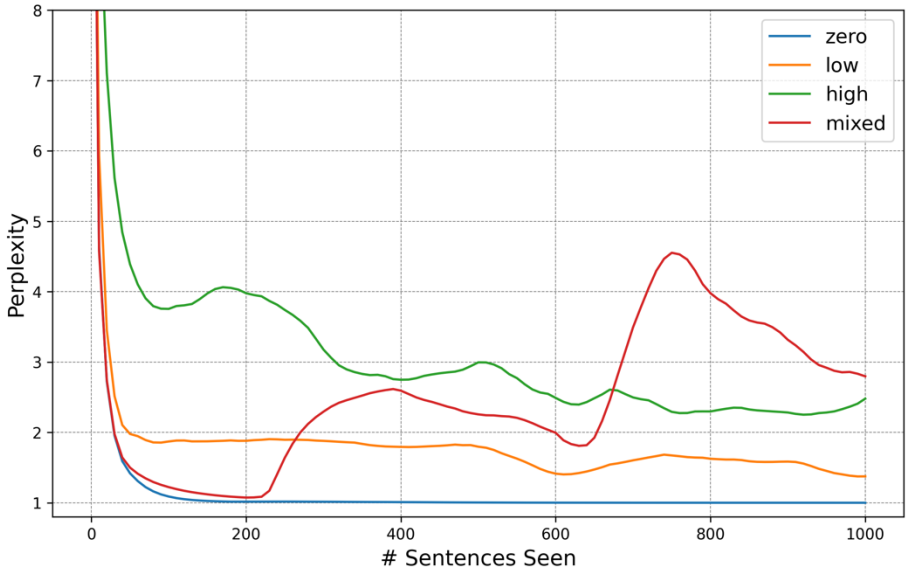


Figure 1. Per-token perplexity in different training scenarios.

[6] It is important to note that the sentences generated for this experiment were pseudo-sentences, occasionally ungrammatical, a factor deemed irrelevant for the purposes of this experiment. The goal was to vary the input systematically to observe the corresponding effects on perplexity, regardless of grammatical accuracy.

As illustrated in Figure 1, by far the largest factor in perplexity reduction is the presence of the training data or its subset in the test sentence, driving the massive decrease of perplexity during the first few batches alone. This well-known and well-understood conclusion should be nuanced, however, as different levels of variability reveal distinct patterns which have important stylometric implications. When the model is fine-tuned without any perturbation, the perplexity on the test sentence rapidly decreases to 1, as the only thing the model needs to do is to predict the single correct token. In scenarios with low and high variability, the perplexity also decreases but at a slower pace, as the model needs to learn broader patterns rather than memorizing a single sentence. Perhaps most interesting is the mixed variability scenario, where the model quickly adapts to the static input but then becomes increasingly perplexed given the rising variability of the training data.

From this perspective, perplexity can be interpreted as a reflection of "discursive resistance" that a given style poses to the language modeling task. This resistance manifests in variability of language use—a literary style allowing for diverse expressions, idioms, and unconventional syntax presents higher resistance to the model's attempt to anticipate what comes next. The fourth scenario ("mixed variability") offers a simulated example of diachronic language change, from more to less regulated, as the evolving style allows for increasing variability in word choices, which visibly perplexes the model. Therefore, in addition to other formal features identified by computational stylometrists (Herrmann et al. [2015]), literary style can be also defined as a force which impacts the branching factor, or the number of plausible paths that a given language sequence can follow. In situations where multiple continuations of a sequence are syntactically and semantically valid, style serves as a filter, tipping the balance in favor of generations that are more aligned with the given stylistic preferences. Since learning a language style is different from memorizing a dataset, this simple experiment also indicates that different styles have different minimum perplexity thresholds which the models might be unable to reduce further, as there could be multiple *equally valid* continuations at each generation step.

This preliminary experiment develops the observations of Miaschi et al. [2021] regarding perplexity's lack of inverse correlation with readability. Casual style (aka "free style") features a higher branching factor than formal style; hand-written characters are more difficult to recognize than printed ones. The most readable sentences are the most variable, where not only different subjects, objects, and predicates are possible but also different ways of expressing and describing them. In such free-style contexts, the probability mass is distributed over multiple sequence continuations, which, as per Shannon's definition, is bound to increase the entropy of the incoming signal. By contrast, the highly technical or formal discourse tends to feature smaller variability and more concentrated probability distribution, which leads to lower perplexity.[7]

3.1.2 Quantifying Maospeak

What happens if we take these insights into account and compare the real texts from our dataset? In the following experiment, I used three open-source Chinese GPT models (Table 3) pre-trained on different amounts and types of data and featuring different number of parameters.[8] I randomly sampled 10,000 sequences of 100 characters from each class in the dataset, preserving the original punctuation. I then used model-specific tokenizers and calculated the

[7] Such a formalist definition of style also implies that style has direct bearing on the *content* of speech, as it might drive the probability of certain words and expressions to 0, effectively eliminating them from language.
[8] The comparative approach is helpful in corroborating the results. The same strategy has been used by Chakraborty et al. [2023] in their search for high-entropy words in de-watermarking experiments.

average per-token perplexity for each dataset for each model. To increase the explainability of the results, I calculated the cross-entropy loss (negative log-likelihood) one sequence at a time and exponentiated the average loss at the end.

Model	Parameters	Training Data	Data Size
IDEA CCNL/ Wenzhong 2.0-GPT2-3.5B-chinese	3.5B	Wudao Corpus	300GB
uer/gpt2-chinese-cluecorpussmall	1.5B	CLUECorpusSmall (news, Wikipedia, comments, blogs)	14GB
ckiplab/gpt2-base-chinese	0.1B	Wikipedia, CNA corpus (news)	4.3B tokens

Table 3. Basic statistics of the three Chinese GPT models used in the experiment.

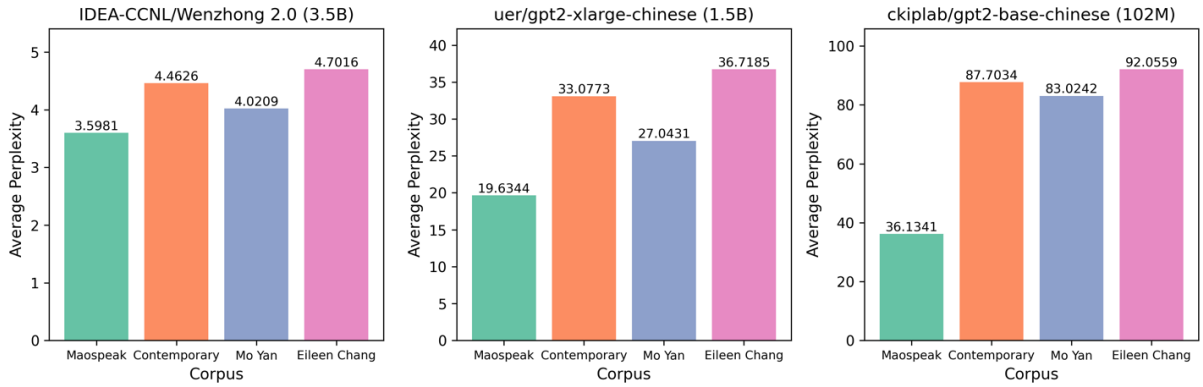


Figure 2. Average per-token perplexity across different Chinese GPT models.

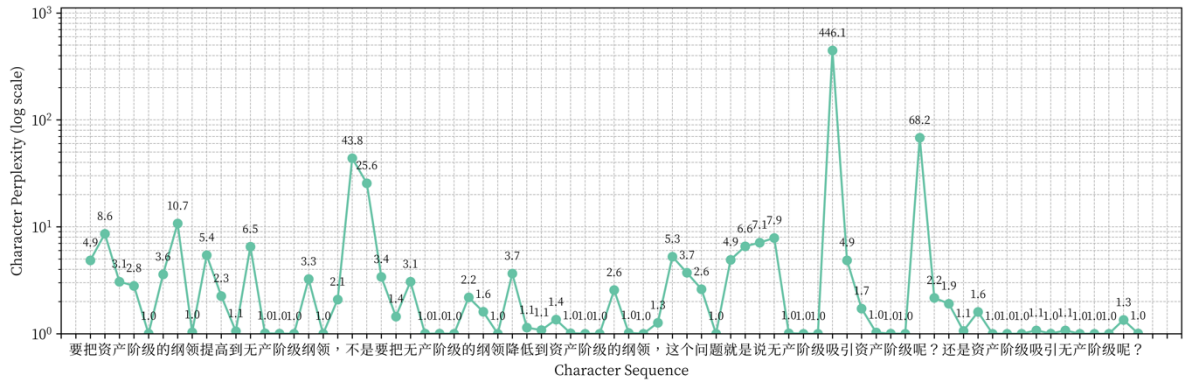
The results (Figure 2) demonstrate that Maospeak features a much lower perplexity than the other three corpora, indicating its higher predictability as far as the models' familiarity with human language is concerned. By contrast, Eileen Chang's writings exhibit the highest per-token perplexity, reflecting more unexpected word choices (again, from the pre-trained models' perspective). The relatively high perplexity of Mo Yan's writing raises a question to consider for his critics, although the results are not conclusive either way. The perplexity scores reflect the abundance of political discourse in the publicly available datasets such as Wikipedia, which accounts for the models' greater familiarity with the Maoist vocabulary. In addition to this common-sense explanation, however, the low perplexity of Maospeak, observable across all three models, provides additional insights into the interrelated features of engineered languages as elaborated by Ji Fengyuan [2003]:

- Highly technical and domain-specific terms and expressions, such as political slogans or administrative verbiage, which allow for little variability in word choices.
- Lack of words and phrases that allow for multiple sequence continuations, notably the quotidian language where the branching factor tends to be the highest.
- Less creativity in word usage (i.e., new metaphors, rare vocabulary, unconventional syntax are mostly absent from Maospeak).

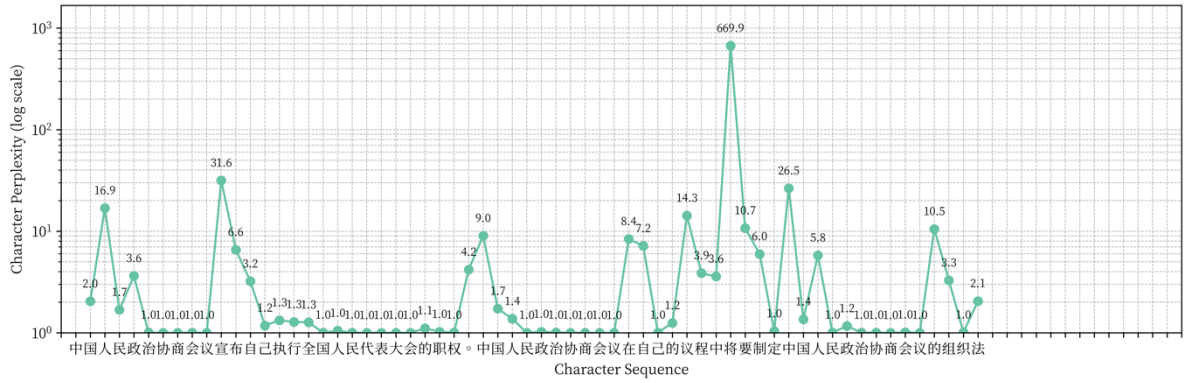
All these characteristics can be observed in the example passages selected from the corpus of Mao Zedong (lowest perplexity) and Eileen Chang (highest perplexity), respectively (Tables 4 and 5). Notice that Eileen Chang switches constantly between simple and rare structures in each sentence: 可以看见 "one could see" (5a), 坐在 "to sit on" (5b), and 去找 "to look for" (5c) are

Low Perplexity (Maospeak)

a) “The goal is to elevate the bourgeoisie program to the level of the proletariat program, not to lower the proletariat program to the level of the bourgeoisie program. The question being posed is: is the proletariat attracting the bourgeoisie, or is the bourgeoisie attracting the proletariat?” (Speech at the Seventh Congress of the Chinese Communist Party, April 1945)



b) “The Chinese People’s Political Consultative Conference declares itself to perform the duties of the National People’s Congress. In its agenda, the Chinese People’s Political Consultative Conference will formulate the organizational law...” (Opening Speech at the First Plenary Session of the Chinese People’s Political Consultative Conference, September 1949)



c) “The class struggle between the proletariat and the bourgeoisie, the class struggle among various political forces, and the ideological class struggle between the proletariat and the bourgeoisie are all long-term, tortuous, and sometimes even very intense.” (“On the Correct Handling of Contradictions Among the People,” February 1957).

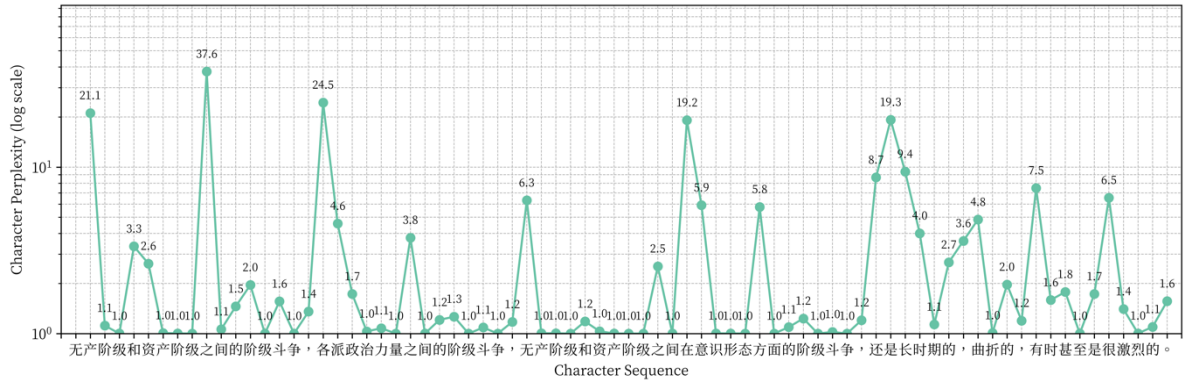
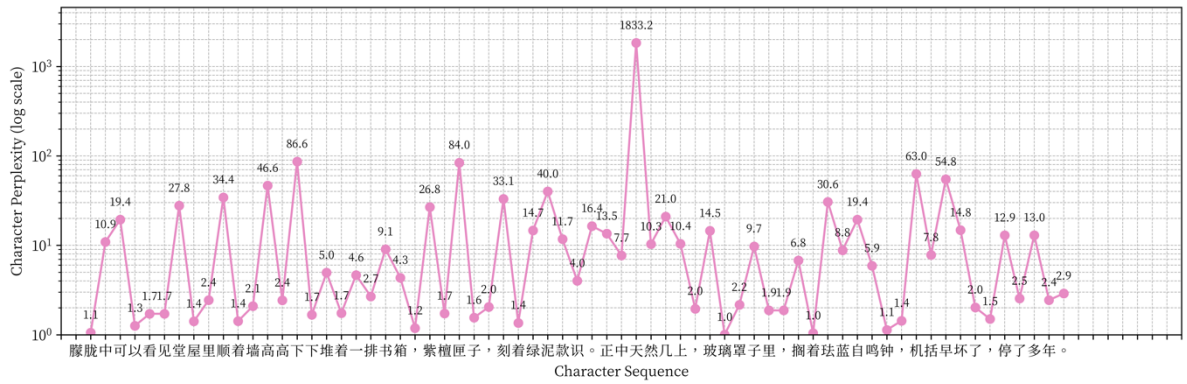


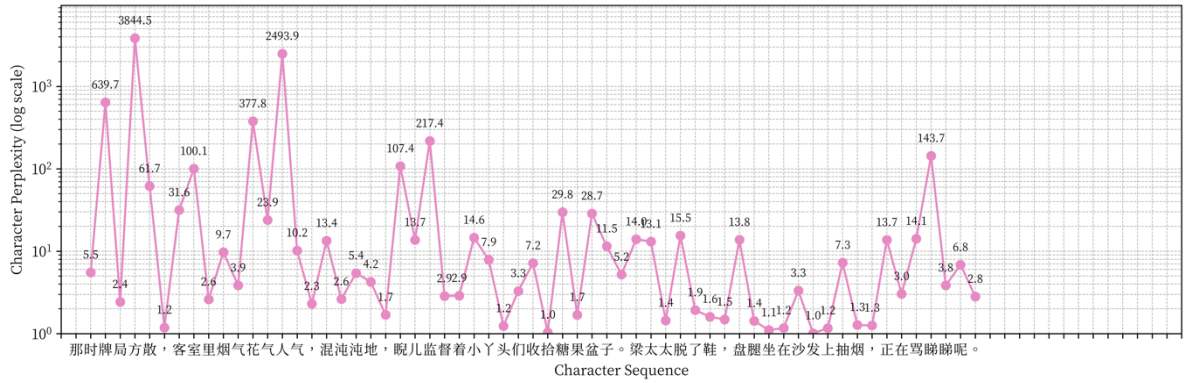
Table 4. Examples of low-perplexity sequences from the Maospeak corpus (Wenzhong 2.0).

High Perplexity (Eileen Chang)

a) “In spite of the gloom, one could see, on the bookshelves that lined the walls, long rows of slipcases made of purplish sandalwood into which formal-script characters had been carved, then painted green. On a plain wooden table in the middle of the room, there was a cloisonne chiming clock with a glass dome over it. The clock was broken; it hadn’t worked in years.” (*Love in a Fallen City*, translated by Karen Kingsbury)



b) “The mah-jongg games had just broken up; the smoke in the living room was a choking, dizzying haze. Glint was supervising some junior maids as they picked up the snack trays. Madame Liang had taken her shoes off, and now sat cross-legged on the sofa with a cigarette, scolding Glance.” (*Love in a Fallen City*, translated by Karen Kingsbury)



c) “Jade Flower had two brothers studying at university and the rooms were turned into guest rooms when they visited. They slept on the teak-frame bedsteads with stretched woven rattan. Julie only occupied one room so she closed the sliding door between them. Later, when she went next door to look for a book to read, she noticed an inkstone and a brush on the desk, and a piece of paper loosely scrunched up.” (*Little Reunions*, translated by Jane Weizhen Pan and Martin Merz).

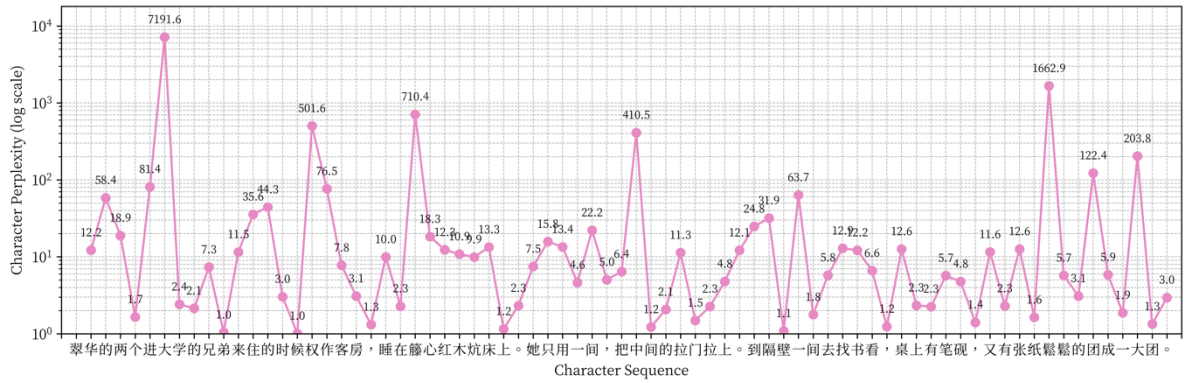


Table 5. Examples of high-perplexity sequences from the Eileen Chang corpus (Wenzhong 2.0).

simple sequences with a high branching factor, featuring verbs with multiple possible objects and adverbials, whereas 籐心红木炕床 “teak-frame bedsteads with stretched woven rattan” (5c) is a rare expression which the model is unlikely to have seen frequently during training. Verbs at phrase boundaries and lists of objects separated by commas tend to be harder to predict, given that the diegetic contexts in which fictional characters appear and act allow for many possible events to take place. This constant mixing of simplicity and rarity literally keeps the model perplexed, which manifests as rapid oscillations and long, high-perplexity ranges in the Eileen Chang graphs (Table 5). By contrast, Maospeak features repetitive and internally redundant token sequences as well as stock phrases such as 全国人民 “the whole nation” (4b) or 中国人民 “the Chinese nation” (4b), which belong to the most common expressions in modern standard Chinese. It also features ideological binaries, notably the 无产阶级 “proletariat” and 资产阶级 “bourgeoisie” (4c), composed of words which almost never appear alone and thus effectively prime the model, decreasing its perplexity. This phenomenon can be observed in the long, low-perplexity valleys in the Maospeak graphs (Table 4), indicating that the model guesses the incoming tokens with next to no difficulty.

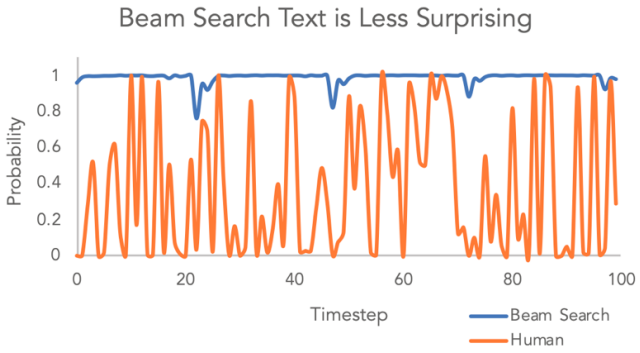


Figure 3. The probability assigned to tokens generated by Beam Search and humans, given the same context. From Holtzman et al. [2019].

The above measurements of perplexity, analyzed in more detail in section IV, allow us to reconnect with the opening discussion. Maospeak resembles “machine text” more than “human text” in the sense elaborated by Holtzman et al. [2019] in their work on neural text degeneration. Specifically, Holtzman and colleagues argue that human-written text is not the most probable text. “Natural language rarely remains in a high probability zone for multiple consecutive time steps, instead veering into lower-probability but more informative tokens” (Holtzman et al. [2019, 7]). The average per-token probability of natural text is much lower than the text generated by likelihood-maximizing text-generation techniques (Figure 3). Beam search, for example, aims to select the most probable sequence of words by evaluating and keeping track of the most promising options at each step of generation. This, in theory, should ensure coherence and fluency, but in practice the maximization-based decoding methods output text that is bland, incoherent, or gets stuck in repetitive loops. As noted by Holtzman et al. [2019], Grice’s maxims of communication, which follow from Shannon’s information theory, hold that people tend to convey information that adds value to the conversation. This can happen only if the incoming linguistic signal cannot be fully anticipated by interlocutors. Maospeak can be thus compared to pre-RLHF text generation techniques which would maximize the probability of the next token and produce redundant, loopy content (Table 4).[9]

[9] RLHF, or Reinforcement Learning from Human Feedback, is a technique which combines machine learning models with human feedback to fine-tune the models’ responses, leading to significant improvements in language generation by aligning model outputs more closely with human preferences.

3.1.3 Perplexity Games

Perplexity can be also used to compare the learnability and generalizability of different language styles. This section introduces Perplexity Games, a comparative framework where different corpora, rather than different models, are pitted against each other in a tournament-style evaluation. Reversing the conventional use of perplexity as a method for model evaluation, the following experiment has three primary objectives: (1) to verify which language style is most easily learnable, as evidenced by the model's post-training perplexity on its own training data; (2) to determine which style endows a model with the most robust and generalizable language understanding, as evidenced by the model's ability to predict texts written in other styles; (3) to identify which style is most difficult to imitate by models trained on other styles. The cross-evaluation of models, each ingrained with the stylistic nuances of its training corpus, seeks to quantify the universality and adaptability of the learned linguistic patterns. The winner of Perplexity Games is the style which strikes a balance between the three general objectives: (1) needs little compute to be mastered; (2) predicts other styles easily; (3) resists prediction by models trained on other styles.

In the following experiment, a 102M-parameters GPT-2 model with randomly initialized weights was trained for exactly one epoch on 5,000 non-overlapping sequences of 100 characters sampled randomly from each class in the dataset. The sequences preserved the original punctuation. This setting guaranteed that each model was given the exact same amount of compute and training data to learn the respective style. After training, the average per-token perplexity with respect to the training corpus as well as the remaining three corpora was calculated. The results of one such round can be visualized as a perplexity heatmap (Figure 4). This process has been repeated 20 times, each time with a different random initialization of model weights and different sampling and ordering of training and test sequences.

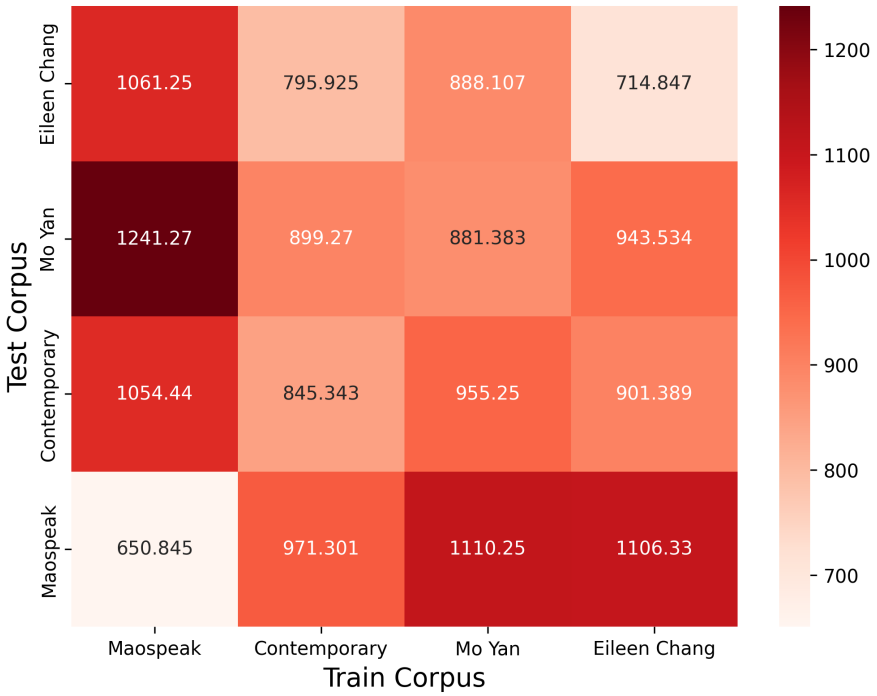


Figure 4. One round of Perplexity Games illustrated as a heatmap. Notice that the lowest values are placed along the ascending diagonal (self-perplexity), but there are exceptions. In particular, the model trained on the Contemporary corpus is less perplexed by the Eileen Chang corpus than by its own training data.

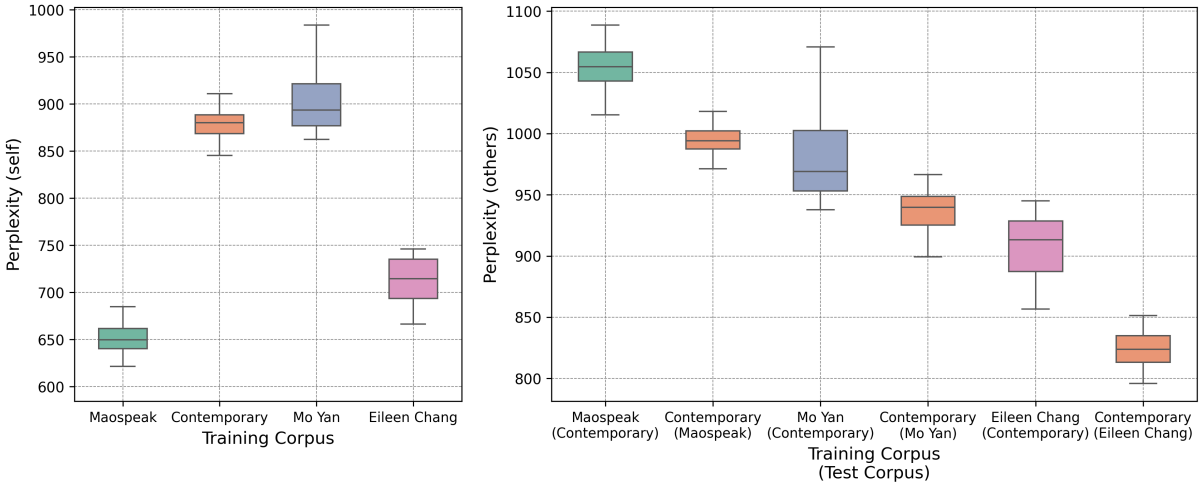


Figure 5. Perplexity Games: learnability (left, lower is easier) and generalizability (right, lower is better) of different language styles, as expressed through perplexity of the model post-training. The right-side graph shows only the comparisons between the Contemporary corpus and other corpora and should be read in pairs: the first column shows perplexity of the Maospeak-trained model on the Contemporary corpus, the second: the Contemporary-trained model on Maospeak, etc.

The Perplexity Games framework provides insights into the stylometric characteristics of different corpora. The Maospeak corpus stands out for its learnability (Figure 5 left), where its low perplexity indicates that GPT models can easily learn and predict its language patterns. The high learnability of Maospeak does not translate to its generalizability, however. For example, the models trained on Maospeak struggle more with the Contemporary corpus than vice-versa, as shown in the right-side graph. On average, Maospeak models perform poorly on texts from other sources, suggesting that Maospeak users may face challenges when adapting to new and unfamiliar environments. This finding agrees with Haiyan Lee's argument about socialist "comradeship" as a mode of sociality that excludes strangers (Lee [2014]) and stands in sharp contrast to the Contemporary corpus, which, while being harder to learn, exhibits much greater flexibility. Having seen various excerpts from Chinese novels, Contemporary models achieve low perplexity when dealing with texts in different styles and win most of the games (Figure 5 right). This adaptability is particularly evident when testing the Contemporary-trained models on the Maospeak corpus, the latter posing a significant challenge to models trained on single-author texts. The versatile and adaptable linguistic structure within the Contemporary corpus carries important implications as regards the role of literature in counteracting ideology-induced cognitive constraints, to be explored in section IV. For now, the above results suggest that reading and writing widely might be the optimal strategy to win the Perplexity Games: doing so enhances not only one's predictive power but also resistance to prediction.

3.2 Entropy

If the previous experiments demonstrated that the open-source GPT models are better at predicting and learning Maospeak than literary texts, an intrinsic experiment can tell us more about the entropy of the corpora themselves. Entropy is a measure of uncertainty: a fair dice has a higher entropy than an unfair one, as the result of each throw is less predictable. Stylometrically speaking, a lower entropy indicates a more "unfair" distribution over the vocabulary, privileging certain terms over the others, whereas a higher entropy entails a more uniform distribution and a greater variety of words and phrases.

In this experiment, I randomly sampled sentences with replacement up to a total of 100,000 characters for each class in the dataset. I then computed the average Shannon entropy for

individual Chinese characters, sequences of characters (*n*-chars), individual words (which might contain more than one Chinese character), and sequences of words (*n*-grams) within each sample. I repeated this process 1,000 times for each corpus to build a distribution of entropy values and calculate the means for each category. The maximum entropy could be theoretically reached if each element (whether *n*-char or *n*-gram) appeared only once in the entire corpus. In this case, the entropy would simplify to $\log_2(number\ of\ unique\ elements)$, which for 5-chars in our 100,000-characters-long samples is $\log_2(100000 - 5 + 1) \approx 16.61$. The minimum entropy, on the other hand, could be achieved in the unlikely case where the whole corpus consisted of just one element repeated 100,000 times, or $\log_2(1) = 0$.

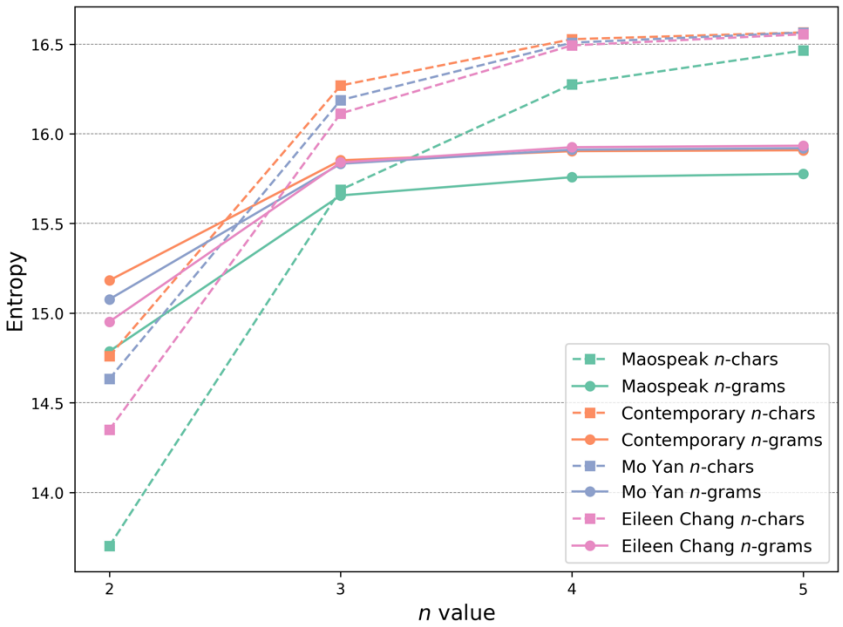


Figure 6. Corpus-specific mean entropies for different *n*-chars and *n*-grams. For larger *n*, most sequences become rare or even unique to the corpus, which flattens the probability distribution and causes entropy to converge to the maximum possible value. Notice, however, that Maospeak converges more slowly than others, given its repeated use of longer sequences and a less uniform distribution.

Figure 6 presents the mean entropies for different values of *n*, showing that the Maospeak corpus consistently exhibits the lowest entropy. This can be attributed to high number of strongly lexicalized *n*-chars present in Maospeak, such as 马克思主义 ("Marxism"), 社会主义 ("socialism"), 小资产阶级 ("petite bourgeoisie"), 帝国主义 ("imperialism"), and 中国共产党 ("the Chinese Communist Party"), as well as political *n*-grams, such as 列宁主义的理论和中国革命的实践 ("Leninism and the Chinese Revolutionary Practice"), 南京国民党反动政府 ("the Reactionary Kuomintang Government in Nanjing"), 坚定正确的政治方向 ("staunch and correct political direction"), etc. Maospeak is more *n*-gramic than literature, relying heavily on fixed, frequently repeated sequences of words or characters. By contrast, the multi-character lexicalized terms and combinations of such terms, while present, are less prominent in literary texts, and those that do appear are less thematically specific: 的时候 ("at the time when"), 到上海来 ("come to Shanghai"), 一个女孩子 ("one girl"), 不知为什么 ("don't know why"), 在我的脑海里 ("in my mind"), etc. Proper names, such as 高密东北乡 ("Gaomi Northeast Township") or 世钧 ("Shijun") are not prominent nor systematic enough to significantly decrease the overall entropy of the corpus.[10]

[10] Mo Yan often sets his stories in the fictional Gaomi Township, which is inspired by his hometown in Shandong Province, China. Shijun is a character from Eileen Chang's 1948 novel *Half a Lifelong Romance* 半生缘.

3.3 Vocabulary

Although an *n*-gram may occur frequently within a given corpus, its commonality across multiple corpora may render it less distinctive or representative of any specific body of texts. This last experiment classifies texts based on the presence or absence of characteristic words and expressions. TF-IDF, short for Term Frequency-Inverse Document Frequency, is a numerical statistic that reflects how important a word is to a document in a corpus. It is defined as follows:

$$\mathrm{TF-IDF}(t, d, D) = \mathrm{TF}(t, d) \cdot \mathrm{IDF}(t, D) \tag{2}$$

Where:

- t is a term (word),
- d is a document containing the term,
- D is the corpus or collection of documents.

The Term Frequency (TF) is calculated as:

$$\mathrm{TF}(t, d) = \frac{\text{Count of term } t \text{ in } d}{\text{Total terms in } d} \tag{3}$$

The Inverse Document Frequency (IDF) is calculated as:

$$\mathrm{IDF}(t, D) = \log\left(\frac{\text{Count of docs in } D}{\text{Count of docs with term } t}\right) \tag{4}$$

If a term t is frequent locally (3) and rare globally (4), its TF-IDF in the given document will be large. TF-IDF emphasizes words that are unique to a particular document while giving less weight to words that are common throughout the entire dataset. Training an explainable classifier, like a Decision Tree or Logistic Regression model, on such terms, enables the identification of features that most strongly indicate a particular style. The features discovered in this way are relational and thus do not tell us anything about any particular style "as such," providing instead a way to distinguish different styles from each other.

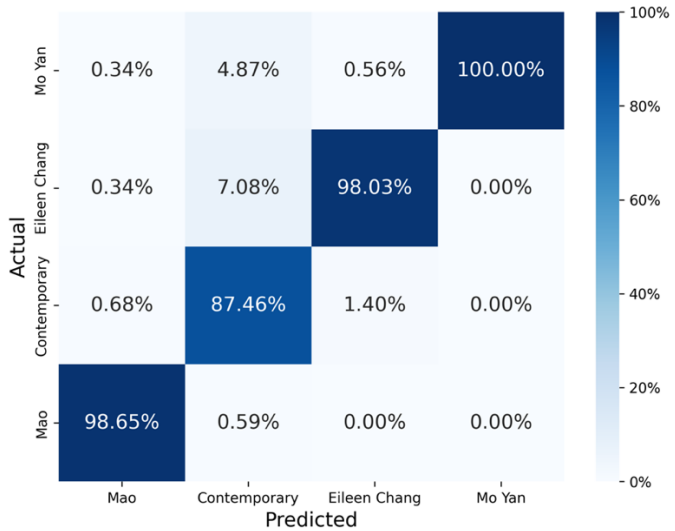


Figure 7. Confusion matrix for the Random Forest classifier, trained with 300 TF-IDF features on 1,000-word fragments, as evaluated on 1,629 test fragments. Values are normalized over predicted conditions (columns).

The dataset in this experiment was constructed by dividing the four corpora into equal-length sequences and then sampling 3,000 sequences from each of them without replacement. In cases where fewer segments were available, I did not oversample. For each class, the sampled fragments were split into training (80%) and test sets (20%). I then trained a Random Forest classifier with 100 decision trees, each with a maximum depth of 15, using the 300 most significant words based on TF-IDF values from the training set as features. These features were computed using the *TfIdfVectorizer* from the *scikit-learn* library for Python. The optimal size of the sequence was determined by evaluating the classifier on the held-out test fragments. The classifier performed well on fragments longer than 500 words, achieving more than 92.5% accuracy, suggesting a high degree of reliability in distinguishing different forms of writing (Figure 7).

To gain deeper insights into which words are most indicative of a particular corpus, SHAP values (*SHapley Additive exPlanations*; Lundberg and Lee [2017]) have been computed post-training. SHAP values allow us to measure the impact (the average marginal contribution) that each feature has on the model's output. They can reveal which words have the strongest influence in classifying a text as coming from the Mao Zedong corpus, for example, essentially pointing out the vocabulary that distinguishes Maoist prose from other styles.

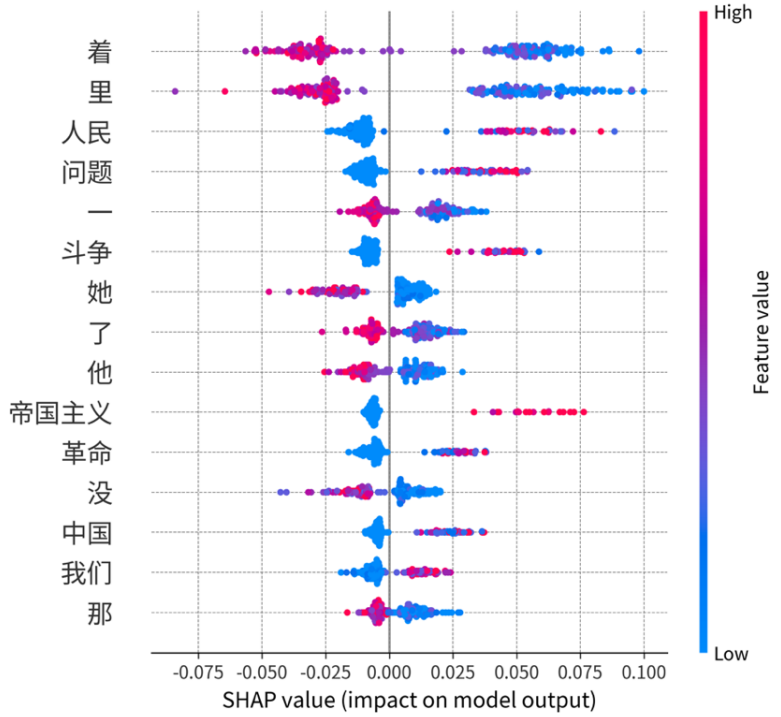


Figure 8. SHAP values for the class "Maospeak" in the Random Forest classifier, computed on 200 test samples (documents). Each row contains 200 dots, and each dot in each row indicates a different test sample (document); the dot's color indicates the TF-IDF value of a given word in the given test sample; horizontal placement indicates the word's contribution (SHAP value) to each individual prediction.

As shown in Figure 8, the computed SHAP values locate the main difference between Maospeak and the literary discourse in the point-of-view markers. Literary texts are characterized by particles such as 着 (continuous state marker), 了 (change of state marker), and 里 ("in") as well as singular pronouns 他 ("he") and 她 ("she"), which ground narratives in the actions and thoughts of depicted characters. This result agrees with the previous work on third-person pronouns and perceptual categories as "strongest indicators of fictionality" (Piper [2017, 17]). By contrast, Mao-style prose speaks on behalf of the first-person plural 我们

("we") and gathers depersonalized, political terms such as 人民 ("the People"), 中国 ("China"), 帝国主义 ("imperialism"), or 斗争 ("struggle"). Crucially, the visualization of SHAP values demonstrates that literary style is not only defined by the features present in a text but also those that are absent, as shown by the horizontally displaced blue dots which strongly contributed (even when absent, i.e., when bringing the TF value to zero or close to zero) to the model's final predictions. The absence of specific words and expressions is the result of the probability distribution unique to each style which reduces the likelihood of their occurrence to 0, as suggested above (section 3.1.1 and note 7).

IV DISCUSSION

4.1 Modern theories of ideology

All the above experiments complement the existing scholarship on ideology, despite significant differences in focus. For example, scholars have long pointed to the ritualistic nature of Maoism, where the cultic texts such as the *Little Red Book* were not intended to transmit new insights but rather to reinforce existing beliefs (Aijmer [1996]; Leese [2011]; Pang [2017]). In ritualistic contexts, the heightened predictability serves to create a sense of unity and shared understanding among followers (Hobson et al. [2018]); the ability to recite a text from memory is a proof of allegiance rather than a typical communicative act. The above results support the argument that one of the primary functions of ideology is the ritualization of language at the expense of its ability to convey information. This is exactly what happened to Maospeak, as the revolutionary leadership lost control over discursive meaning during the most violent years of the Cultural Revolution, when opposing Red Guard factions appealed to the exact same Mao quotations to struggle against each other (Ji [2003, 139-150]).

On the other hand, while traditional analyses of ideological discourse emphasize the division between "us" and "them" (Dijk [1998, 2022]), the computational framework focuses on distributional characteristics of texts. Thus, rather than foregrounding narrative strategies of negative other-presentation and positive self-presentation, I focus on the statistical limitations that a particular discourse sets on the possibilities of expression. This approach does not contradict political theories: political polarization can be seen as a special case of entropy reduction, restricting the range of language used for depicting oneself and others. Louis Althusser's concept of Ideological State Apparatuses (ISAs) supports this view by showing how linguistic patterns can serve as ISAs that enforce identity through a limited linguistic repertoire (Althusser [1971]), while Terry Eagleton's perspective on ideology as naturalization further clarifies why and how perplexity is minimized in ideological contexts (Eagleton [1991]). As these and other thinkers have noticed, ideology actively interpellates individuals through discourse and affect. In what follows, I build upon but also move beyond these traditional accounts to argue that the discovered statistical limitations not only reflect the ideological constraints imposed on linguistic creativity but also suggest deeper cognitive processes at work which have made such constraints possible in the first place.

4.2 Predictive Processing

One of the major theories in cognitive science suggests that human cognition is governed by predictive processing (Clark [2013]; Orlandi and Lee [2019]; Caucheteux [2023]). The brain constructs and continually updates a hierarchical representation of the world to decrease entropy of incoming signals and reduce discrepancies between expected and actual sensory inputs (Radulescu et al. [2020]; Schmid [2023]). Due to the high metabolic costs associated with

maintaining a detailed perceptual representation of the environment, cognitive systems must strike a balance between being precise and being efficient. This goal can be achieved through hierarchical pattern extraction: utilizing established frameworks to infer information top-down rather than re-analyzing every piece of incoming data from the ground up. Inputs that align with predictive models reinforce existing hierarchical representations and accelerate processing, while misestimated inputs might prompt adjustments to these models, ensuring that cognition remains both adaptive and efficient. This concept is vividly illustrated in reading, where the predictability of words significantly impacts processing speed. The graded pre-activation of likely next words simplifies early lexical processing and influences eye movements (Cutter et al. [2020]). Readers fixate more quickly on words that are predictable from the preceding context than on those that are not; predictable words are also skipped more frequently and elicit smaller N400 responses, indicating less cognitive effort required to process them (Kutas and Hillyard [1984]; Liu et al. [2020]; Slattery and Yates [2018]). Recent research suggests that activations in the middle layers of the Transformer models resemble activation patterns in the human brain during word prediction tasks, and this resemblance increases with the model's prediction accuracy (Caucheteux and King [2022]).

While predictive processing is essential for our survival, this hierarchical framework, with its built-in stabilization mechanisms, can be easily hijacked by external forces. Wheeler et al. [2020], for example, see ideologies as high-level predictive models built on top of the motoric and perception models. Ideologies underlie the formation of shared beliefs and values within a group, as individuals collectively minimize prediction errors in their understanding of and acting upon the social world. This shared error minimization can facilitate cooperation and predictability within the group but also lead to the reinforcement of biases and the polarization of beliefs. Kitto and Boschetti [2013] provide a geometric interpretation of the same process: given a number of agents operating within local and global decision frames (represented as orthogonal vectors "yes/no") and characterized by varying levels of "conformity" (the urgency to bring their local frame as close as possible to the global frame) and "consistency" (the urgency to align their decisions with their local frame), the authors show that over time the agents tend to self-organize towards alignment within groups (Figure 9); the entropy of the system, which is the sum total of binary entropies of individual agents, is expected to decrease when alignment increases.

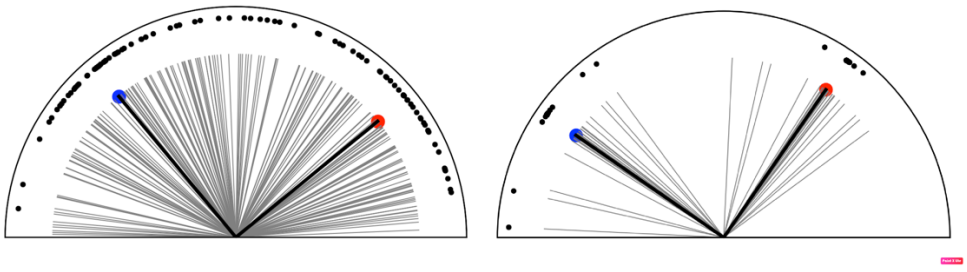


Figure 9: Multi-agent decision making dynamics, illustrating the initial, high-entropy state (left) and the post-100 iterations, low-entropy state (right). The agents' local frames (thin orthogonal vectors) align over time around the global frame (thick orthogonal vectors), showcasing the system's tendency towards ideological convergence. Reproduced and adapted from Kitto and Boschetti [2013].

These and other studies suggest that political systems may have emerged to narrow the potential dynamics of communal life. In highly ideological contexts, the feedback loop minimizes the cross-entropy between the instantiations of the dominant discourse in the form of officially sanctioned texts, speeches, movies, etc., and the patterns (local frames) internalized by

individuals who inhabit such contexts. As there are relatively few options to choose from at any position in the incoming symbolic data, the mental model does not need to consider the probabilities of other, less likely symbols, thus saving the metabolic energy. This increased predictive accuracy in turn skews the learned distribution towards an ever-smaller subset of possible "tokens." This process ultimately leads to overfitting, as the model increased specialization results in impaired ability to generalize and adapt to new contexts.

4.3 The Role of Literature and the Arts

While such feedback loops are inherent in any dynamic system, they can be moderated through periodic expansions of the realm of possibilities, which are bound to cause temporary increases in the entropy of the symbolic parser, even as this parser continues to reorganize itself with the view of attaining the optimal state (Figure 10). To use the phrasing from Kitto and Boschetti [2013], such irregular "resets" can be understood as perturbations of local frames in some of the agents. The intermittent introductions and integrations of new cognitive dimensions not only uncover new differences and similarities within the environmental data but also reframe existing contrasts and parallels, preventing the system from settling down.

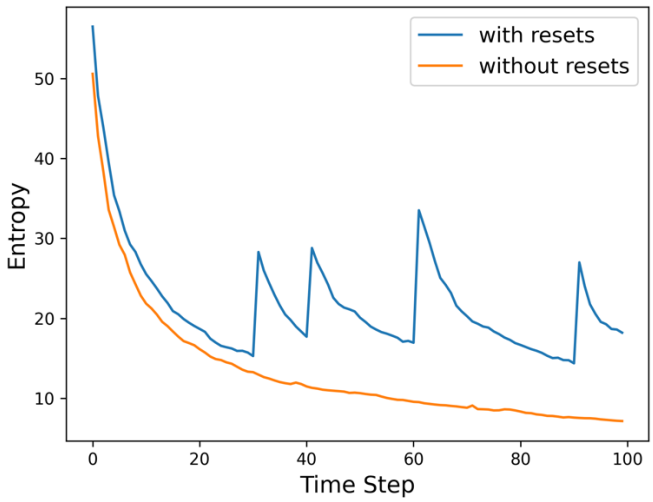


Figure 10. Evolution of entropy over time in two different scenarios (with resets and without resets), adapted from Kitto and Boschetti [2013].

It is from this perspective that fiction reading and humanistic education become especially important. As shown above, language models tend to find narrative texts harder to predict than political discourse. Insofar as the Bayesian model is valid for both artificial and natural intelligence, increasing one's exposure to various language data counters the influences of ideologies on our linguistically mediated perceptions of the world and increases the perplexity of our imaginations. While the benevolent influence of literature is by no means guaranteed, with various "authoritarian" or "foundational" fictions and *romans à clef* aiming to reduce social entropy (Suleiman [1992]; Sommer [1991]; Anderson [1991]), good literature introduces opaque personalities which cannot be fully explained within a single cognitive frame. Encounters with characters that resist single interpretations might improve mind-reading skills (Zunshine [2006]; Bernini [2016]), provoke reflective changes (Nussbaum [1990]), and lead to greater misalignment between the global and local frames. Moreover, the strong focus on subjective experience in literary texts, manifested formally through the pronoun-rich vocabulary, might hone the habit to meta-represent information, i.e., to ascribe a particular claim to a specific person/entity, instead of allowing it to circulate freely in the space of

knowledge (Zunshine [2006, 47-73]; Cosmides and Tooby [2000]). Without this ability, individuals may perceive their own thoughts as originating from a source outside of their own intentions or perceive subjective opinions as unquestionable truths. Metarepresentation reverses this process and enables us to recast global frames as local ones.

V DISCUSSION

In this article, I have conducted a series of experiments to analyze the stylistic properties of Maospeak, an engineered language which reinforces the political tenets of Maoism through its distributional characteristics. I have argued that ideologies function as likelihood-maximization text-generation techniques which reduce the scope of possible sequence continuations and hijack the predictive mechanisms of human cognition. Ideological discourse is more *n*-gramic, featuring long, strongly lexicalized phrases which do not allow for variability in their immediate linguistic contexts. While the foregoing analysis applies chiefly to Maoism, the presented findings are also pertinent to more recent phenomena, in China and beyond (Barmé [2012]). Discussions centered around "diversity," for example, though laudable in principle, might become monolithic if genuinely diverse viewpoints are excluded and the discourse ossifies into a set of politically correct, low-perplexity phrases. Another example is auto-completion in smartphone keyboards: while timesaving, this formidable technology reduces our ability to write creatively and reinforces select patterns of thought. Yet another example is the institutionally supported elimination of local languages in the interest of standardized speech.

The above experiments also show that the conventional use of perplexity as a standardized metric for model evaluation overlooks its stylometric potential. In Maospeak, sentence construction tends to narrow the field of possible continuations, creating a linguistic path that allows only a limited set of subsequent tokens. This linguistic constraint is emblematic of a "closed time" in a narrative text, a concept closely related to Gary Saul Morson's notions of "sideshadowing" and "open time" (Morson [1994, 2013]). Morson builds on Mikhail Bakhtin's theory of the polyphonic novel to argue that the various techniques of sideshadowing found in the literary works of Chekhov, Tolstoy, and Dostoevsky allow the narrative and life to become isomorphic, or similar to each other; in such stories, the "temporality of lived experience triumphs over the temporality of completed structure," and "possibilities are in excess of actualities" (Morson [1994, 41, 83], Bakhtin [1984]). By contrast, utopian narratives present a closed sequence of events that are anisomorphic with respect to life and feature a predetermined, "foreshadowed" conclusion. From this perspective, engineered languages such as Maospeak embody closed time by funneling the narrative down predetermined paths, leaving little room for the contingent and the unexpected. This reduced variability in word sequences corresponds to lower perplexity and higher learnability. Conversely, literary texts embrace "open time," where every word is sideshadowed by many other possible, even if ultimately not chosen words.

There are many likely objections to the present approach. One is that comparing different genres, such as political pamphlets and literary works, is unjustified. Such critique, however, inadvertently underscores the very argument made by Ji Fengyuan in her study on linguistic engineering in the PRC [2003]. Human language inherently oscillates between modes that resemble the characteristics of political pamphlets and those of literary artifacts. The comparison between genres serves not to unfairly juxtapose dissimilar categories but to illuminate the wide spectrum of language use and its capacity to embody elements of both. Not everything that Mao Zedong said and wrote belongs to Maospeak, nor is literature free from ideological influences, as literary texts might be perplexing at relatively low levels of language

comprehension (word choices, e.g.), while reinforcing select patterns of thought at higher levels (arguments and ideas). This article has focused on word-level predictability; future work could expand the paradigm of Perplexity Games to deeper layers and longer timescales.

Another likely critique will come from the creative users of language models. In his book *My Life as an Artificial Creative Intelligence*, for example, Mark Amerika celebrates the advent of generative AI as a new source of creative potential. The language artist produces prompts and the language model responds with counter-prompts, the two converging into a "hybridized form of interdependent consciousness" (Amerika [2022, 6]) which, supposedly by virtue of its de-subjectivized automatism, builds immunity to the "parasitical attack of neoliberal capitalism's propaganda machine" (Amerika [2022, 59]). While I agree that the "fuzzy digressions" of GPTs can help us break the writer's block, Amerika's uncritical embrace of the human-machine auto-completion leaves undertheorized the main objective that drives the bulk of contemporary text generation: producing texts that are likely. The most popular loss functions used in language model training are designed to punish artificial intelligence for not aligning with the *status quo*, rather than rewarding it for being different from it. The so-called "lingual spontaneity" (Amerika [2022, 27-63]) might be a mere storm in a teacup of fore-computed conclusions.

It should be noted, finally, that the present contribution is inherently comparative; it labels a given language style as "closed" only in relation to more "open" styles, both defined in terms of their own entropy and the dynamic relationship between their representative linguistic artifacts and the human subject. There is no specific information threshold that distinctly separates ideology from non-ideology, nor is there a vantage point from which all ideologies could be identified. Such comparative character acknowledges the flexible boundary between the human and the machine. As Chakraborty et al. [2023] puts it, "the text produced by newer LLMs is nearly indistinguishable from human-written text from a statistical perspective." Language models of the future might either turn into a friendly reminder of what it means to write and speak like a human or force us to innovate beyond their ever-improving predictive capabilities. Either way, Perplexity Games are bound to continue.

ACKNOWLEDGEMENTS

The author would like to thank the organizers of the 3rd International Conference on Natural Language Processing for Digital Humanities (NLP4DH), which took place at Waseda University on December 1-3, 2023, as well as the anonymous reviewers for their useful comments and questions. Special thanks are due to Aaron Gilkison and Dr. Heidi Huang for their feedback on the earlier versions of this manuscript. Any remaining errors are my own.

REFERENCES

Aijmer, G. Political Ritual: Aspects of the Mao cult during the Cultural "Revolution." *Theory, Culture & Society*. 1996;11(2-3):135–154.

Althusser, L. *Ideology and Ideological State Apparatuses. Notes towards an investigation.* Lenin and Philosophy and Other Essays*,* Monthly Review Press, 1971.

Amerika, M. *My Life as an Artificial Creative Intelligence*. Stanford University Press (Stanford, CA), 2022.

Anderson, B. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso (London, New York), 1991.

Bakhtin, M. *Problems of Dostoevsky's Poetics*. University of Minnesota Press, 1984.

Barmé, G. R. "New China Newspeak" 新华文体. *China Heritage*, 2012. URL https://chinaheritage.net/journal/on-new-china-newspeak/ Accessed: 25/09/2023.

Bernini, M. The opacity of fictional minds: transparency, interpretive cognition and the exceptionality thesis. In *The Cognitive Humanities: Embodied Mind in Literature and Culture,* pages 35–54. Palgrave Macmillan (London), 2016.

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J., and Teich, E. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, 2020.
Burrows, J. F. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 1987;2(2):61–70.
Cameron, D. *Verbal Hygiene. Essential Business Psychology*. Routledge (London), 1995.
Caucheteux, C., Gramfort, A., and King, JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour* 2023;7:430–441.
Caucheteux, C., Gramfort, A., and King, JR. Deep language algorithms predict semantic comprehension from brain activity. *Nature Scientific Reports* 2022;12:16327.
Caucheteux, C., King, JR. Brains and algorithms partially converge in natural language processing. *Communications Biology* 2022;5(134).
Chakraborty, M., Tonmoy, S. M. T. I., Zaman, S. M. M., Sharma, K., Barman, N. R., Gupta, C., Gautam, S., Kumar, T., Jain, V., Chadha, A., Sheth, A. P., and Das, A. Counter-turing test ct2: Ai-generated text detection is not as easy as you may think – introducing ai detectability index, October 2023. URL https://arxiv.org/abs/2310.05030
Chang, C., S. Nastase, and Hasson U. Information flow across the cortical timescale hierarchy during narrative construction. In *Proceedings of the National Academy of Sciences (PNAS)* 119, No. 51, 2022.
Chomsky, N. The false promise of ChatGPT. *The New York Times*, 08/03/2023.
Clark, A. Whatever Next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 2013;36(3):181–204.
Cohen, T. and Pakhomov, S. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, 2020.
Cosmides, L. and Tooby, J. Consider the source: The evolution of adaptations for decoupling and metarepresentation. In *Metarepresentations: A Multidisciplinary Perspective*, pages 53–115. Oxford University Press (Oxford), 2000.
Cutter, M. G., Martin, A. E., and Sturt, P. The activation of contextually predictable words in syntactically illegal positions. *Quarterly Journal of Experimental Psychology*, 2020;73(9):1423–1430.
Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding, October 2018. URL https://arxiv.org/abs/1810.04805
Dijk, T. A. v. *Ideology: A Multidisciplinary Approach*. SAGE Publications Ltd, 1998.
Dijk, T. A. v. Ideology in cognition and discourse. In *Mapping Ideology in Discourse Studies*, pages 137–156. De Gruyter Mouton (Berlin, Boston), 2022.
Eagleton, T. *Ideology: An Introduction*. Verso (London), 1991.
Epstein, M. Relativistic patterns in totalitarian thinking: an inquiry into the language of soviet ideology. Occasional Paper 243, The Woodrow Wilson International Center for Scholars, Kennan Institute for Advanced Russian Studies (Washington, D.C), 1991.
Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 2017;32:4–16.
Fabien, M., Villatoro-Tello, E., Motlicek, P., and Parida, S. BertAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON),* pages 127–137, 2020.
Herrmann, J. B., van Dalen-Oskam, K., and Schöch, C. Revisiting style, a key concept in literary studies. *Journal of Literary Theory*, 2015;9(1):25-52.
Hobson, N. M., Schroeder, J., Risen, J. L., Xygalatas, D., and Inzlicht, M. The psychology of rituals: An integrative review and process-based framework. *Perspectives on Psychological Science*, 2018;13(3):323–349.
Holtzman, A., Buys, J., Forbes, M., and Choi, Y. The curious case of neural text degeneration, April 2019. URL https://arxiv.org/abs/1904.09751
Huang, W., Murakami, A., and Grieve, J. ALMs: Authorial language models for authorship attribution, January 2024. URL https://arxiv.org/abs/2401.12005
Huang, X., Deng, L., and Acero, A. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.
Ji, F. *Linguistic Engineering: Language and Politics in Mao's China*. University of Hawaii Press, 2003.
Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling, Februrary 2016. URL https://arxiv.org/abs/1602.02410
Kitto, K. and Boschetti, F. Attitudes, ideologies and self-organization: Information load minimization in multi-agent decision making. *Advances in Complex Systems*, 2013;16(02n03):1–37.
Klemperer, V. *The Language of the Third Reich: LTI, Lingua Tertii Imperii: A Philologist's Notebook*. Continuum, 2006. Originally published in German as *LTI – Lingua Tertii Imperii: Notizbuch eines Philologen* in 1947.
Kodner, J., Payne, S., and Heinz, J. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023), August 2023. URL https://arxiv.org/abs/2308.03228
Kutas, M. and Hillyard, S. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 1984;307(5947):161-163.
Laughlin, C. What Mo Yan's Detractors Get Wrong. *ChinaFile*, 11/12/2012. URL https://www.chinafile.com/what-mo-yans-detractors-get-wrong. Accessed 25/09/2023.
Lee, H. *The Stranger and the Chinese Moral Imagination.* Stanford University Press, 2014.
Leese, D. *Mao Cult: Rhetoric and Ritual in China's Cultural Revolution*. Cambridge University Press, 2011.

Li, T. 李陀. 汪曾祺与现代汉语写作——兼谈毛文体 [Wang Zengqi and modern Chinese writing: Also on the style of Mao's writings]. 花城 [Huacheng], 1998.
Link, P. Politics and the Chinese Language: What Mo Yan's Defenders Get Wrong. *Asia Society*, 27/12/2012. URL https://asiasociety.org/blog/asia/politics-and-chinese-language-what-mo-yans-defenders-get-wrong. Accessed 25/09/2023.
Link, P. *An Anatomy of Chinese: Rhythm, Metaphor, Politics*. Harvard University Press, 2013.
Liu, Z., Liu, X., Tong, W., and Fu, F. Word's contextual predictability and its character frequency effects in Chinese reading: Evidence from eye movements. *Frontiers in Psychology* 11, 2020.
Lundberg, S., and Lee, S. A Unified Approach to Interpreting Model Predictions, May 2017. URL https://arxiv.org/abs/1705.07874
Lutosławski, W. Principes de stylométrie. *Revue des études grecques*, 1898;41:61–81.
McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., and Anderson, A. Detecting individual decision-making style: Exploring behavioral stylometry in chess, August 2022. URL https://arxiv.org/abs/2208.01366
Miaschi, A., Alzetta, C., Brunato, D., Dell'Orletta, F., and Venturi, G. Is neural language model perplexity related to readability? In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it,* 2020.
Miaschi, A., Brunato, D., Dell'Orletta, F., and Venturi, G. What makes my model perplexed? A linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures,* 2021.
Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. Detect-GPT: Zero-shot machine-generated text detection using probability curvature, January 2023. URL https://arxiv.org/abs/2301.11305
Morson, G. S. *Narrative and Freedom: The Shadows of Time.* Yale University Press (New Haven), 1994.
Morson, G. S. *Prosaics and Other Provocations: Empathy, Open Time, and the Novel*. Ars Rossica. Academic Studies Press (Boston, MA), 2013.
Mosteller, F. and Wallace, D. L. Inference in an authorship problem. *Journal of the American Statistical Association*, 1963;58(302):275–309.
Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR),* 2017;50(6):1-36.
Nikolaev, D., Karidi, T., Kenneth, N., Mitnik, V., Saeboe, L., and Abend, O. Morphosyntactic predictability of translationese. *Linguistics Vanguard,* 2020;6(1).
Nussbaum, M. C. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford University Press (New York), 1990.
Orlandi, N. and Lee, G. How Radical Is Predictive Processing? In *Andy Clark and His Critics*. Oxford University Press, 2019, 206-221.
Ou, W., Ding, S. H. H., Tian, Y., and Song, L. Scs-gan: Learning functionality-agnostic stylometric representations for source code authorship verification. *IEEE Transactions on Software Engineering*, 2023;49:1426–1442.
Pang, L. *The Art of Cloning: Creative Production during China's Cultural Revolution*. Verso (London and New York), 2017.
Patel, A., Rao, D., Kothary, A., McKeown, K., and Callison-Burch, C. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, 2023.
Piantadosi, S. Modern language models refute Chomsky's approach to language, 2023. URL https://lingbuzz.net/lingbuzz/007180
Pichel Campos, J. R., Gamallo, P., and Alegria, I. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155, 2018.
Piper, Andrew. Fictionality. *Journal of Cultural Analytics* 2016;2(2): 131-142.
Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
Radulescu, S., Wijnen, F., & Avrutin, S. Patterns bit by bit. An entropy model for rule induction. *Language Learning and Development*, 2020;16(2):109–140.
Rivera-Soto, R. A., Miano, O. E., Ordonez, J., Chen, B. Y., Khan, A., Bishop, M., and Andrews, N. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, 2021.
Ruder, S., Ghaffari, P., and Breslin, J. G. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution, September 2016. URL https://arxiv.org/abs/1609.06686
Rybicki, J., Eder, M., and Hoover, D. L. Computational stylistics and text analysis. In *Doing Digital Humanities*. Routledge, 2016, 123-144.
Sari, Y., Stevenson, M., and Vlachos, A. Topic or style? Exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, 2018.
Schoenhals, M. Demonising discourse in Mao Zedong's China: People vs non-people. *Totalitarian Movements and Political Religions*, 2007;8(3):465–482.
Schmid, S., Saddy, D., Franck, J. Finding hierarchical structure in binary sequences: Evidence from Lindenmayer grammar learning," *Cognitive Science*, 2023;47(1):e13242.
Seroussi, Y., Zukerman, I., and Bohnert, F. Authorship attribution with topic models. *Computational Linguistics*, 2014;40(2):269–310.
Shannon, C. E. Prediction and entropy of printed English. *The Bell System Technical Journal*, 1951;30(1):50–64.

Slattery, T. J. and Yates, M. (2018). Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology (Hove),* 2018;71(1):250– 259.
Sommer, D. *Foundational Fictions: The National Romances of Latin America*. Duke University Press, 1991.
Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology,* 2009;60(3):538– 556.
Suleiman, S. *Authoritarian Fictions: The Ideological Novel as a Literary Genre*. Columbia University Press, 1992.
Sun, A. The Diseased Language of Mo Yan. *The Kenyon Review*, Fall 2012. URL https://kenyonreview.org/kr-online-issue/2012-fall/selections/anna-sun-656342/ Accessed 25/09/2023.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, June 2017. URL https://arxiv.org/abs/1706.03762
Wang, A., Aggazzotti, C., Kotula, R., Soto, R. R., Bishop, M., and Andrews, N. Can authorship representation learning capture stylistic features? In *Transactions of the Association for Computational Linguistics*, 2023;11:1416–1431.
Wang, X. *Modernity with a Cold War Face: Reimagining the Nation in Chinese Literature across the 1949 Divide*. Harvard University Press, 2013.
Wheeler, N. E., Allidina, S., Long, E. U., Schneider, S. P., Haas, I. J., and Cunningham, W. A. Ideology and predictive processing: coordination, bias, and polarization in socially constrained error minimization. *Current Opinion in Behavioral Sciences*, 2020;34:192–198.
Wu, K., Pang, L., Shen, H., Cheng, X., and Chua, T.-S. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, 2023.
Zunshine, L. *Why We Read Fiction: Theory of Mind and the Novel*. Ohio State University Press, 2006.