Shaping Collective Memory

An Analysis of Communication Styles Used in Memoirs of the Paris Commune

Mariella Daghfal, Kaede Johnson, & Shayan Khajehnouri
EPFL Course DH-412

I. Introduction & Research Question

We investigate the relationship between memoirs, collective memory, and the Paris Commune, a revolutionary uprising that took place in Paris in spring 1871. Both participants in and opponents to the uprising published their own narrative accounts of the Commune after it was violently put down by forces of the new French Republic. Because Commune supporters wrote to counter the narratives circulated by the French government, communication style was essential in shaping the collective memory of the wider public.

We collect several memoirs and testimonies written by Commune supporters and opponents, invoke the concepts of collective memory, defamiliariation, and literary language's effect on the portrayal of history, and implement computational tools - LSA, LDA, TF-IDF, token extraction, and a metric to measure linguistic "strangeness" - to compare communication styles used by Commune memoirists and, ultimately, understand how these communication styles crafted separate collective memories. More precisely, our main research question is the following:

How do memoirist communication styles shape collective memory among supporters and opponents of the Paris Commune?

II. Secondary Literature

The Commune of Paris was a revolutionary socialist government that existed between March 18 and May 28 of 1871. Emerging months after the fall of the Second French Empire during the Franco-Prussian War, its anti-establishment founding coalition was composed of workers and leftist groups keen on infusing the body governing Parisians with more democratic principals. Agents of the Commune - dubbed *Communards* along with general supporters of the uprising - introduced free public education, workers' cooperatives, and the separation of church and state during their brief time in power, and believed their defiance of the French Republic was morally and economically just. Conservative forces, meanwhile, were appalled by the Commune's actions, and the new Republican government temporarily headquartered in Versailles had little patience for an armed, populist movement occupying its capital city. In May 1871, French government forces - dubbed the *Versaillais* - bombarded Paris for weeks before launching a military offensive inside the city which culminated in a week-long battle known as the "Bloody Week." Those prominent *Communards* who subsequently failed to escape France were executed or imprisoned for their revolutionary activity. [1] [2]

Such diametrically opposed camps are bound to yield different historical memories on the significance, justification, and events of March, April, and May 1871 - memories we can witness today through contemporary writings on the subject. To understand how pro- and anti-Commune memories formed and persisted following the Commune's fall, we study the output of memoirists who presented their own narrative of the Commune to the public, paying particular attention to how these narratives contributed to collective memory.

"Collective Memory" was first coined by the sociologist Maurice Halbwachs in 1925 [3]. According to Halbwachs, rather than being an isolated, personal experience, memory is influenced by social structures, cultural norms, and historical processes that rely on those with whom an individual interacts. Memory is therefore a dynamic process responsive to the social situations in which it is are rooted. It is also a form of social training and, as such, leads to a collective experience of memory. Indeed, Halbwachs believed collective memory is an essential means through which communities comprehend themselves and the past.

In his paper *Three facets of collective memory*, psychologist Henry Roediger underlines the importance collective memory plays in shaping a person's individual identity and draws attention to the distinction between collective memory and formal history [4]. He considers collective memory to be history as people remember it, which subsumes or supplants the reality of formal history. Collective memory can form among any group, whether small (family, school) or large (political party, nation). Given our focus on the Commune and, specifically, authors writing about the Commune, we consider how our two ideological camps' literary representatives contribute to their respective collective memories.

One of our chief analytical approaches is to investigate how literary flexibility might afford memoirs more efficacy in constructing memory. In answer to this question, historian Jennifer Jenson Wallach claims memoirs can provide a "more detailed description of historical reality than that contained in any other single historical context," thanks to the "vehicle of literary art. [5]" She does not claim all literary techniques are capable of effecting this detailed description, but rather those which ask the reader to pause, think, and "re-feel" - symbolism, irony, metaphor, and allegory, in her view. That memoirists "essentially transform themselves into characters," also provides insight into how historical figures conceived of society and the arbitrariness of life that standard historians tend to erase by imposing structure. By studying the usage of figurative and non-figurative sentences in our corpus, we

can arrive at the specific memories Commune memoirists emphasize as well as the lens with which they conceive of pro- or anti-Commune agents in general.

Much of Wallach's argument is ultimately derived from the concept of "defamiliarization" coined by Russian formalist Viktor Shklovsky [6]. According to Shklovsky, defamiliarization occurs when language challenges readers' habitual perception of the world, stimulates their engagement with the text, and encourages them to grapple with preconceptions. Literary theorist Terry Eagleton goes on to claim this "'estranging' or 'defamiliarizing' effect... forces [readers] into a dramatic awareness of langauge," which, "paradoxically, brings [them] into a fuller, more intimate possession of experience. [7]" Familiar language suggests a familiar reality, fashioning a crude historical experience readers can readily digest rather than something more challenging yet more accurate. Eagleton therefore presents one view of literary language "as a set of deviations from a norm." Putting aside the question of the "norm" for a moment, it follows from Shklovsky, Eagleton, and Wallach that quantifying linguistic deviations from said norm - the "strangeness" of a Commune memoirist's language, in other words - is one way we can glean insight into literary language's power over collective memory formation.

Existing qualitative research involving stylistic analysis of Commune memoirs is limited. Sidonie Verhaeghe, with her paper *Les victimes furent sans nom et sans nombre: Louise Michel et la mémoire des morts de la Commune de Paris*, analyzes figures of speech used to describe the scale of death during the "Bloody week" [8]. The exact number of deaths is debated, leading Commune supporters like Louise Michel to overestimate the scale of the *Versaillais'* killings, Commune opponents to underestimate it, and a fault line to form between the collective memories of distinct cohorts. It is this sort of analysis which we hope to emulate with digital methods and extend to our memoirs overall. Accordingly, our research focuses on extracting word distributions, diction, figures of speech, and "strange" language in order to analyze pro- and anti-Commune literary communication styles and, ultimately, the implications for collective memory formation.

III. Data

The dataset we analyze is housed in Gallica, digital counterpart to the Bibliothèque Nationale de France [9] [10]. It is composed of twenty-four memoirs and testimonies. Twelve of these memoirs were written by opponents of the Commune. The other twelve were written by exiled supporters of the Commune. The organization of these memoirs into "pro-Commune" and "anti-Commune" factions is based on Gallica's description of the memoirists and verified by our own reading of portions of the memoirs.

Some of these memoirs were published soon after the fall of the Commune (such as Vergès d'Esboeufs' *La Vérité sur le gouvernement de la Défense nationale, la Commune et les Versaillais*, André Léo's *La Guerre sociale*, and Gustave Lefrançais's *Étude sur le mouvement communaliste à Paris en 1871*). Their authors, then in exile, were keen on publishing their version of the facts, especially because a growing number of publications hostile to the Commune were appearing after June 1871 and trials were simultaneously taking place against prominent *Communards*. Other pro-Commune memoirs were published much later, with dates ranging from 1877 to 1909; the drive to prevent suppression of the *Communard* perspective lived on - as did the drive to keep the Commune alive in the public's collective memory.

The full list of memoirs we analyze is available in Appendix Table I.

IV. Methods

A. Pre-Processing

Gallica maintains a scanned version of each memoir as a PDF file and .txt versions of those PDFs to which it has applied optical character recognition (OCR) tools - meaning tools to convert images of text into text files. We download all available .txt versions of our corpus' works for analysis. For those memoirs Gallica has not converted into .txt files using OCR tools, we download the aforementioned PDF versions and create .txt files using OCRmyPDF, an open source OCR toolkit [11]. [1] Gallica provides estimates for the share of words that have been properly transcribed via OCR in a given .txt file; for our corpus, most estimates range between 93% and 98%. Still, serious transcription errors persist.

After collecting all twenty-four memoirs as .txt files, we remove leading and trailing text from each narrative, correct OCR mistakes common to our corpus (for example, letters in *Il* are often mistaken for punctuation), remove abbreviations, page headers, and page numbers, and clean date references in those memoirs which delineate content based on date. We also use Python's SpellChecker package to correct tokens that would autocorrect to words incorporated into our analysis; these words include *commune, versaillais, communards, adolphe thiers, théophile ferré,* and *ivre*.

We identify sentence breaks using a punctuation-finding regular expression. For LSA, LDA, and the "Strangeness" embedding metric described in Subsection B below, we lemmatize all tokens and drop tokens that are non-alphabetic, less than two characters in length, or stop words. Stop words are supplied by Python's nltk library [12]. Unless otherwise noted, we use Python's spaCy library [13] to implement all other nlp operations.

B. Analytical Methods

We first apply LSA to our memoirs following the process described by Stephen DeCanio [14]. That is, we represent each memoir as a set of tokens, weight each token according to log entropy, and approximate memoir representations using the second and third primary components of SVD on a term-memoir matrix. The first primary component is ignored due to its proclivity to target memoir length.

Further analysis often involves our sentence extraction tool. This tool splits memoirs using the aforementioned sentence

[1] Documentation available at https://ocrmypdf.readthedocs.io/en/latest/index.html

breaks and retains only those sentences containing a given input token. We note that issues in OCR transcription may lead to some sentences being overlooked, cut short, or inappropriately combined with preceding or succeeding sentences.

The sentence extraction tool is first used to implement LDA (Latent Dirichlet Allocation) and TF-IDF (Term Frequency - Inverse Document Frequency) methods on sentences from our corpus, mainly to "give color" to the patterns found with LSA. For TF-IDF, we select a word or phrase based on our historical understanding of the Commune and extract all sentences from all memoirs containing this word or phrase. Separating the sentences according to the bias of their author, we apply TF-IDF weighting to the tokens in each group of extracted sentences, uncovering the words that are largely unique to a given perspective's writing on a shared topic. Our LDA implementation relies on a similar structure - save for the fact that extracted sentences are **not** separated according to bias of the author. Instead, we apply LDA to all extracted sentences for a given word or phrase, recover the set of suggested topics, and inspect these topics manually to determine whether they are more closely associated with pro- or anti-Commune memoirists.

Finally, after investigating the presence of figurative language surrounding one particularly arresting word discovered through close reading, we adapt a metric developed by Mao, Lin, and Guerin to quantify what we term the "strangeness" of a given sentence [15]. First, for a given token i within sentence s, we calculate the average word embedding vector across all tokens $k \in s$ s.t. $k \neq i$ using FastText's Wikipedia word embeddings [16]. Call this average $\overline{v}_{sans_i}$. We then develop a set of tokens C_i containing i and all synonyms and hypernyms for i listed in WordNet [17] using nltk and identify the token in C_i closest to $\overline{v}_{sans_i}$ using cosine similarity. Call this token bc. The "strangeness" of a token i, then, is quantified with the cosine similarity between the embeddings of i and bc ($cs_{i,bc}$); intuitively, the lower $cs_{i,bc}$, the "stranger" it is for i **rather than** bc to appear among i's neighboring tokens. Collecting all sentences written by Joanni d'Arsac, Edmond de Goncourt, and Élie Reclus - the three memoirists in our corpus who structure their narrative around the daily or near-daily events of the Commune - we find the minimum $cs_{i,bc}$ per sentence and the average of such minimums per week by author. We note that our "strangeness" metric ignores all tokens without embeddings; all such tokens which appear more than three times can be found in Appendix Table II (note that all non-alphabetic tokens are excluded from this metric). Finally, we acknowledge that using Wikipedia embeddings to capture what Eagleton would call a "norm" for 19th-century language is problematic, and so propose an alternative "norm" in our Conclusion section.

V. Results & Interpretation

A similarity matrix for memoir representations approximated with the second and third primary components of SVD - our LSA result - is shown in Figure 1. Comparing anti-to-anti comparisons in the bottom right to pro-to-pro comparisons

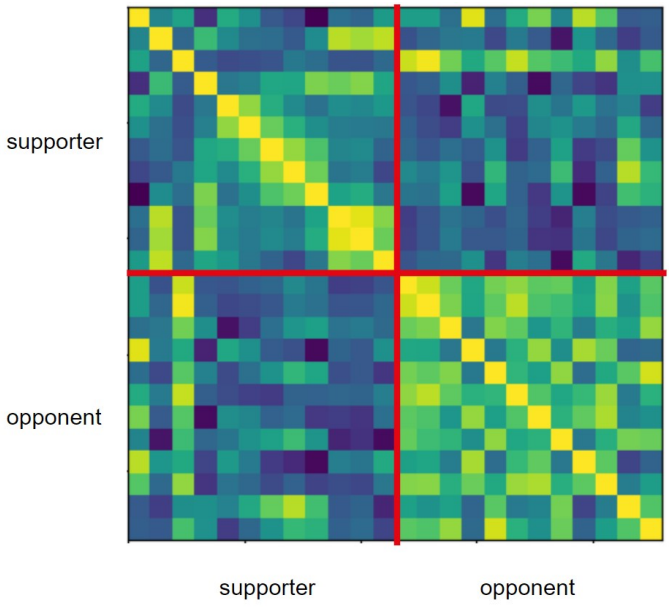

Fig. 1. LSA results (brighter = higher cosine similarity between memoirs' word distributions). Opponents of the Commune show greater cohesion in word usage than do supporters of the Commune. See Appendix Figure 1 for similar matrix but with select memoirs split into multiple parts.

in the top left, we find there is more cohesion in word distribution variation among anti-Commune memoirists than among pro-Commune memoirists (this result weakens but still holds until about half of all non-first-component variation is included in the approximation). In other words, anti-Commune memoirists are more likely to use similar words at similar rates than pro-Commune memoirists. Furthermore, these shared anti-Commune distributions are largely distinct from pro-Commune memoirists' distributions, as the top right or bottom left sections of the heatmap - pro-to-anti comparisons - report low similarities for all but two pro-Commune memoirists.

How do we interpret these results? First, at least in our corpus, critical accounts of the Commune likely coalesce around narratives and tones that are more similar to each other than accounts crafted by Commune supporters, leading to a more standardized collective memory. One intuitive explanation of this phenomenon is that anti-Commune memoirists shared the perspective of the State and did not operate in exile; closer to their historical setting, their ideological colleagues, and an "official" narrative they might readily accept, it would have been easier to access and emulate narratives with which they agreed. Nonetheless, this standardization does not necessarily imply a stronger collective memory, for if pro-Commune memoirs contain more literary language, then Wallach and Eagleton's ideas suggest greater ability to shape memory. Indeed, greater incidence of literary language would lead to more distinct word distributions and, consequently, a pro-to-pro comparison square like that seen in Figure 1. Lending credence to this interpretation is a version of the LSA similarity matrix with select memoirs split into multiple parts (see Appendix Figure 1); strong intra-memoir similarities suggest authors' writing

styles motivate similarity matrix patterns in addition to memoir content.

To understand why pro- and anti-Commune memoirists produced different word distributions despite addressing the same historical events, we turn next to TF-IDF and LDA. Table I displays the words both important and unique to a given memoirist cohort (that is, words that surface from TF-IDF weighting) among sentences involving *communards*, *versaillais*, and *Thiers* (surname to the head of the French government during the Paris Commune). We select these words to understand how each memoirist cohort conceived of themselves and of the "opposing" force in the narratives they craft.

TABLE I
TF-IDF RESULTS

Words	**"Pro-Commune" Memoirs**	**"Anti-Commune" Memoirs**
Adolphe Thiers	Prussians/Massacre /Revolution	Collection/Objects/Bronzes
Communards	Prison/Prisoners/Convicted	Column/Army/Barricade
Versaillais	Prussians/Blood /Adolphe Thiers	Shouted/Signal/Boy

Pro-Commune memoirists often link the *versasillais* to the French head of state Thiers, the Prussians (with whom the *Versaillais* formed an alliance to retake Paris), and with blood - a narrative emphasizing the enemy coalition and violence. Anti-Commune memoirists, meanwhile, associate the *Versaillais* with words that have no clear connection or significance. As for the term *communard*, pro-Commune memoirists associate it with the judgements fellow *Communards* experienced at the end of the Commune, while anti-Commune memoirists make dry references to battle ("army", "barricade") or infrastructure ("column" refers to the destruction of the Vendôme Column, a symbolic action which was an affront to the French government).

One interpretation is that when considering the words unique to each cohort, pro-Commune memoirists define both their own political cohort and the opposing cohort with more narrative flourish than do anti-Commune memoirists. To them, the enemy is colluding with a foreign power to bring about the macabre while *Communards* are persecuted. Meanwhile, the Commune's opponents fail to paint a strong, distinct picture of the *versailles* and consider the *Communards* primarily through a tactical or physical lens.

This is further supported by our cohorts' treatment of important individuals. While pro-Commune memoirists uniquely associate Adolphe Thiers with Prussians, massacres, and revolution, anti-Commune memoirists admire his impressive collection of bronzes. Meanwhile, LDA results for sentences involving Théophile Ferré, a *Communard* elected to the Commune council and member of the Sûreté générale commission, yield three coherent topics, from which we sample words more prevalent in either memoirist cohort and display in Table II.

The first topic relates to Ferré's arrest and sentencing to death. Unsurprisingly, sentences linking Ferré to words in this topic are overwhelmingly found in pro-Commune memoirs (trial: 73%, judge: 100%). Terms in the second topic, which relate more to his time heading the police force in the Commune, are much more evenly divided between the two cohorts, though they favor anti-Commune memoirs (police: 63%). Here, the anti-Commune authors emphasize the facts of Ferré's role, while pro-Commune authors emphasize the facts of his dramatic end.

TABLE II
LDA RESULTS

Words	**"Pro-Commune" Memoirs**	**"Anti-Commune" Memoirs**
Theophile Ferré	Judge/Trial /Member/Recognize	Do/Police/Order /Citizen/"Sûreté"

The clear presence of unique, dramatic diction among pro-Commune memoirists' perspectives on universal terms suggests the memory they craft may be more closely linked with ideology even if word distributions are less cohesive across their cohort. To be clear, we do not claim that anti-Commune memoirs are bereft of ideology. Rather, these results suggest anti-Commune memoirists are not weaving their ideology into the narrative in a way that deviates from the events recounted by pro-Commune authors - nor, by extension, in a way that allows the collective memory they craft to stand out.

While LSA, TF-IDF, and LDA revealed a general difference in word distributions between our two cohorts, we did want to investigate how memoirists could share word distributions despite differences in ideology. To do so, we perform a case study on usages of the French word *ivre*.

As shown in Figure 2, pro- and anti-Commune memorists both use *ivre* literally most of the time. However, literal usage still reveals differences in how memoirists conceived of the Commune. Unless referencing a named individual, literal usages of *ivre* tend to characterize the 'opposing' force (*Garde Nationale* in the case of Commune opponents, *Versaillais* in the case of Commune supporters). Opposing usages of *ivre* extend into the figurative realm as well: for pro-Commune memoirists, the *Communards* are "*ivre de vérité* (drunk with truth), [18]" while to opponents they are "*ivre de vengeance et de fureur* (drunk with vengeance and fury). [19]" Diction alone does not shape collective memory. Rather, it is diction informed by intent.

Figure 2 also shows that pro-Commune memoirists use *ivre* in a figurative sense more than twice as often as anti-Commune memoirists. "*Sa besogne terminée,*" says Allemane, "*on vit le bourreau, aussi pâle qu'était pâle la face émergeant du son, quitter son horrible machine et s'éloigner, tel un homme ivre.* [20]" With this simile, Allemane compares an executioner to a drunk man, encouraging the reader to "re-feel" the execution of Allemane's fellow prisoner. A count of all comparison phrases reveals 50% more such phrases in pro-Commune memoirs (in terms of relative token frequency), with

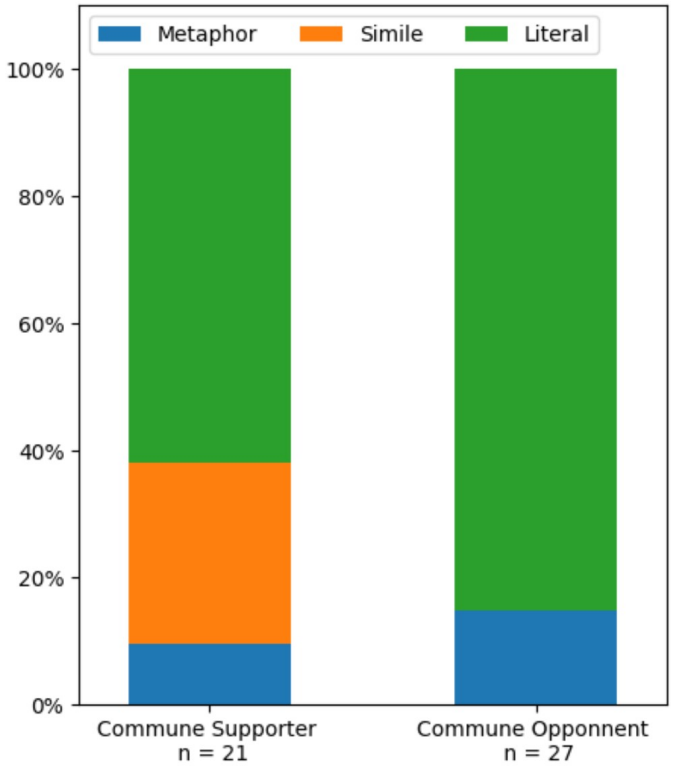

Fig. 2. Classification of *ivre* usages.

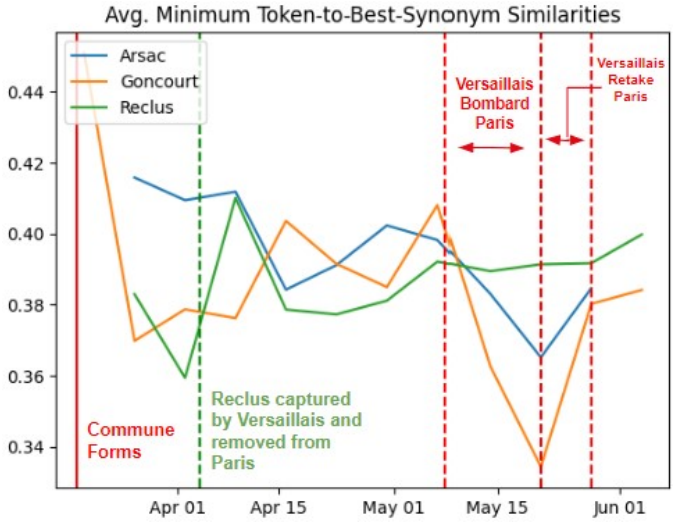

Fig. 3. "Strangeness" metric (lower = sentences use less common word combinations on average). Simultaneous dips for Arsac and Goncourt in the Commune's final weeks suggest these two anti-Commune memoirists, through the paradox of literary discourse posited by Terry Eagleton, impress the same critical events into the mind of their readers (intentionally or otherwise). Reclus' steadiness in the same timeframe suggests his memoir does not will these events into collective memory as forcefully - possibly because he had been captured and removed from Paris long before, rendering him less of a "character" in his narrative.

relatively even distribution across memoirs.[2] Additionally, close reading of a 100-observation random sample from both pro- and anti-Commune memoirs reveals comparison phrases in pro-Commune memoirs are twice as likely to be used in a figurative sense. These admittedly crude calculations suggest that, as with *ivre*, pro-Commune memoirists are more likely to employ similes to 'make literary' passages of their text - a habit which might will more individual moments into the collective memory of the pro-Commune space.

However, as Wallach notes, not all literary techniques impart the strongest historical reality. Rather, it is mostly those which make language "strange" for the reader. Figure 3 plots our "strangeness" metric for sentences written by two anti-Commune memoirists (Goncourt and Arsac) and one pro-Commune memoirist (Reclus) on specific dates during the Commune. Note that based on how our metric is constructed, **lower** values correspond to **less familiar** word combinations.

Simultaneous drops for Goncourt and Arsac between May 8 and May 28 suggest these two anti-Commune authors' writing is more likely to "de-familiarize" during this section of their narrative, pushing events from the *Versaillais'* bombardment of Paris into memory of the Commune on at least two fronts.[3] For Reclus, meanwhile, language strangeness remains relatively steady during the final month of the Commune. This may be due to his capture by *Versaillais* forces and internment thereafter; though he still writes about events occurring in Paris, his absence renders him less of a character in his own account (as Wallach would describe it), therefore robbing him of experience and incentive conducive to effecting a literary style.

Keeping in mind that daily or near-daily language strangeness is only available for three authors in our corpus, we are hesitant to use it to make claims about pro- and anti-Commune memoirists as a whole. Nonetheless, we believe Figure 3 provides a compelling response to our research question: a memoirist's writing becomes more literary or strange in response to flashpoints he or she experiences, and when multiple such "de-familiarizing" turns of language coincide among a cohort of, for example, anti-Commune memoirists, these flashpoints can be willed into a collective memory. This is what occurred for Arsac and Goncourt, and it may be what occurred in other retellings of the Commune by both pro- or anti-Commune memoirists. That the strangeness of Reclus' writing peaks (meaning the metric plotted in Figure 3 reaches a global low) during his dangerous time with the *Garde Nationale* supports this interpretation. In a sense, his capture marks his transition from a memoirist to a more traditional historian, in so far as he begins to describe events he did not personally experience. The emphasis Wallach places on the memoirist as a character in his or her own narrative, then, appears well founded.

[2]Comparison words or phrases: *comme, tel, pareil à, semblable à, de même que, avoir l'air de, faire l'effet de,* and *ressembler à.*

[3]Some examples of sentences identified as "strange" include: (1) "*...aux allées, qu'on dirait macadamisées avec des éclats d'obus, tant il en est tombé, tant il en tombe tous les jours.*" (2) "*...le bleu du ciel ensoleïllé, l'éclosion, la formation, le grossissement lent de nuages, semblales à ces nuages de féerie, d'où sort un génie ou une fée, habillée de papier d'or...*" [21]

VI. Conclusion & Future Steps

Using digital tools to evaluate word distributions, diction, and figurative language usage across a corpus of twenty-four memoirs about the Paris Commune, we find that while anti-Commune memoirists craft a more cohesive collective memory through their writing, pro-Commune memoirists craft a more vibrant and arresting collective memory through greater usage of figurative, unique, and narratively-inclined language. Additionally, our analysis of linguistic "strangeness" possibly reveals the mechanism through which memoir writing pushes specific events into collective memory: collective deviation from linguistic norms that is exacerbated by personal exposure to a singular event.

While these results stand up to several robustness tests performed with our corpus, it is important to note the limits of this project. First, twenty-four memoirs cannot represent the entirety of contemporary writing by authors exposed to the Commune's events. Even if these memoirs received outsized attention from the public upon release, there would have been many more books on the subject, not to mention newspapers and letters. Additionally, OCR errors mean the text we study is not a perfect representation of the text we intend to study. Future efforts should expand the number of memoirs studied, incorporate newspapers and letters provided they are recounting events with the possibility for literary flourish, and hunt down remaining OCR errors.

Implicit in the construction of our "strangeness" metric is the notion that word embeddings trained on French Wikipedia entries represent the linguistic "norm." This is a dubious claim for a corpus of texts written between 1871 and 1909. A more appropriate "norm" might be embeddings trained exclusively on French non-fiction from the late 19th century. Along with expanding the number of memoirs which clearly delineate their content based on date, training the word embeddings on a more appropriate corpus would improve the utility of our strangeness metric.

Finally, we note that authors might employ defamiliarization techniques beyond what we target. Authors might experiment with unconventional storytelling techniques, non-linear timelines, and fragmented narratives. Simile, metaphor, allegory, and imagery identification is an open research topic in NLP; our strangeness metric is only a proxy for these methods, and a rough proxy at that. Future researchers could develop specialized tools to extract literary techniques from our corpus, achieve a more holistic evaluation of literary language presence, isolate the events each memoirist discusses with defamiliarizing language, and observe intersections to map a very clear pipeline of collective memory construction.

References

[1] Alistair Horne. *The Fall of Paris: The Siege and the Commune, 1870-71*. St. Martin's Press, 1966.

[2] Samuel Bernstein. The paris commune. *Science Society*, 5(2):117–147, 1941.

[3] Nicolas Russell. Collective memory before and after halbwachs. *The French Review*, 79(4):792–804, 2006.

[4] Henry L. Roediger. Three facets of collective memory. *American Psychologist*, 76(9):1388–1400, 2021.

[5] Jennifer Jensen Wallach. *Remembering Jim Crow: The literary memoir as historical source material*. University of Massachusetts Amherst, 2004.

[6] Daniel P. Gunn. Making art strange: A commentary on defamiliarization. *The Georgia Review*, 38(1):25–33, 1984.

[7] Terry Eagleton. *Literary theory: An introduction*. John Wiley & Sons, 2011.

[8] Sidonie Verhaeghe. nbsp;les victimes furent sans nom et sans nom-brenbsp;. louise michel et la mémoire des morts de la commune de paris. *Mots*, (100):31–42, 2012.

[9] Gallica - témoignages de communards, https://gallica.bnf.fr/html/und/histoire/temoignages-de-communards.

[10] Gallica - témoignages d'anti-communards, https://gallica.bnf.fr/html/und/histoire/temoignages-danti-communards.

[11] James A. Barlow. Ocrmypdf. https://github.com/ocrmypdf/OCRmyPDF, 2023.

[12] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[13] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[14] Stephen J DeCanio. Ai recognition of differences among book-length texts. *AI & SOCIETY*, 35:135–146, 2020.

[15] Rui Mao, Chenghua Lin, and Frank Guerin. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL), 2018.

[16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[17] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[18] Arnould Arthur. *Histoire populaire et parlementaire de la commune de Paris*. 1878.

[19] Johanni d'Arsac. *La guerre civile et la Commune de Paris en 1871: suite au Mémorial du siège de Paris / par J. d'Arsac*. F. Curot, 1871.

[20] Jean Allemane. *Mémoires d'un communard: des barricades au bagne*. 1906.

[21] Edmond de Goncourt. *Journal des Goncourt: mémoires de la vie littéraire, deuxième série, premier volume*. G. Charpentier et E. Fasquelle, 1890.

[22] Charles Beslay. *La vérité sur la Commune*. 1878.

[23] Victorine Brocher. *Souvenirs d'une morte vivante*. 1909.

[24] Gustave-Paul Cluseret. *Mémoires du général Cluseret*. 1887.

[25] Gaston Da Costa. *La Commune vécue*. 1903.

[26] Vergès d'Esboeufs. *La vérité sur le gouvernement de la Défense nationale, la Commune et les Versaillais*. 1871.

[27] François Jourde. *Souvenirs d'un membre de la Commune*. 1877.

[28] Gustave Lefrançais. *Étude sur le mouvement communaliste à Paris en 1871*. 1871.

[29] Prosper-Olivier Lissagaray. *Les huit journées de mai derrière les barricades*. 1871.

[30] André Léo. *La Guerre Sociale*. 1871.

[31] Élie Reclus. *La Commune de Paris au jour le jour*. 1908.

[32] Paul Bourde. *Les membres de la Commune et du Comité central / par Paul Delion*. A. Lemerre, 1871.

[33] Abraham Sébastien Crozes. *Le capitaine fédéré Révol (Deuxième édition) / récit historique de l'abbé Crozes*. 1872.

[34] Maxime Du Camp. *Les convulsions de Paris. Épisodes de la Commune / Maxime Du Camp*. Hachette et Cie, 1879.

[35] Maxime Du Camp. *Les convulsions de Paris. Les prisons pendant la Commune*. Hachette et Cie, 1879.

[36] Léonce Dupont. *Souvenirs de Versailles pendant la Commune*. E. Dentu, 1881.

[37] Arthème Fayard. *Histoire de la Commune de 1871*. 1871.

[38] Eugène Hennebert. *Guerre des communeux de Paris: 18 mars-28 mai 1871 / par un officier supérieur de l'armée de Versailles*. 1871.

[39] Catulle Mendès. *Les 73 journées de la Commune (du 18 mars au 29 mai 1871)*. Lachaud, 1871.

[40] Henry Morel. *Le pilori des communeux: biographie des membres de la Commune, leurs antécédents, leurs moeurs, leur caractère: révélations*. 1871.
[41] Edgar Rodrigues. *Le carnaval rouge*. 1872.

APPENDIX

Bias	Name	Year	Citation
Pro	Jean Allemane	1906	[20]
Pro	Arthur Arnould	1878	[18]
Pro	Charles Beslay	1878	[22]
Pro	Victorine Brocher	1909	[23]
Pro	Gustave-Paul Cluseret	1887	[24]
Pro	Gaston Da Costa	1903	[25]
Pro	Vergès d'Esboeufs	1871	[26]
Pro	François Jourde	1877	[27]
Pro	Gustave Lefrançais	1871	[28]
Pro	Prosper-Olivier Lissagaray	1871	[29]
Pro	André Léo	1871	[30]
Pro	Élie Reclus	1908	[31]
Anti	Paul Bourde	1871	[32]
Anti	Johanni d'Arsac	1871	[19]
Anti	Abraham Sébastien Crozes	1871	[33]
Anti	Maxime Du Camp	1879	[34][35]
Anti	Léonce Dupont	1887	[36]
Anti	Arthème Fayard	1871	[37]
Anti	Eugène Hennebert	1871	[38]
Anti	Catulle Mendès	1871	[39]
Anti	Henry Morel	1871	[40]
Anti	Edgar Rodrigues	1872	[41]
Anti	Edmond De Goncourt	1890	[21]

TABLE I
FULL CORPUS OF MEMOIRS. ACCESSED VIA GALLICA.

Token	**Frequency**
dévoûment	9
Monlrouge	8
eùt	7
ayec	7
ofliciers	6
etles	6
boureois	6
natioaux	6
MontValérien	6
remercîments	6
diflicile	5
gouverement	5
Prusiens	5
luiême	5
çais	5
êlre	5
veuent	4
noueau	4
natioale	4
ilest	4
mouveent	4
Îls	4
tousles	4
œuerre	4
seuleent	4
pouoir	4
ments	4
privilége	4
Châlillon	4
vingtquatre	4
prêlres	4
Dicêtre	4
qyi	4
lssy	4
Châtilîon	4
Térifocq	4
vaincrez	4

TABLE II
TOKENS WHICH LACK EMBEDDINGS (AND ARE THEREFORE EXCLUDED) IN "STRANGENESS" METRIC

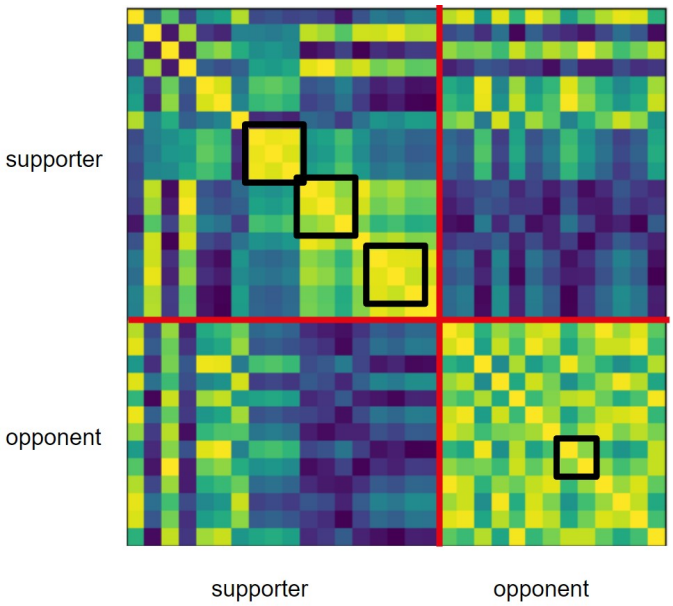

Fig. 1. LSA results when select memoirs are split into multiple parts; similarity matrix for approximated memoir (or memoir subset) representations using the second and third primary components of SVD on log-entropy weighted word occurrences (brighter = higher cosine similarity between memoirs' word distributions; black boxes surround comparisons of subsets of the same memoir). As in Figure 1, opponents of the Commune show greater cohesion in word usage than do supporters of the Commune. Similarity in comparisons of subsets of the same memoir with other memoirs suggest these patterns capture author communication style.