Formalizing Style in Personal Narratives

Gustave Cortal[1,2], Alain Finkel[1],

[1]Université Paris-Saclay, CNRS, ENS Paris-Saclay, LMF, 91190, Gif-sur-Yvette, France
[2]Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France

{gustave.cortal, alain.finkel}@ens-paris-saclay.fr

Abstract

Personal narratives are stories authors construct to make meaning of their experiences. Style, the distinctive way authors use language to express themselves, is fundamental to how these narratives convey subjective experiences. Yet there is a lack of a formal framework for systematically analyzing these stylistic choices. We present a novel approach that formalizes style in personal narratives as patterns in the linguistic choices authors make when communicating subjective experiences. Our framework integrates three domains: functional linguistics establishes language as a system of meaningful choices, computer science provides methods for automatically extracting and analyzing sequential patterns, and these patterns are linked to psychological observations. Using language models, we automatically extract linguistic features such as processes, participants, and circumstances. We apply our framework to hundreds of dream narratives, including a case study on a war veteran with post-traumatic stress disorder. Analysis of his narratives uncovers distinctive patterns, particularly how verbal processes dominate over mental ones, illustrating the relationship between linguistic choices and psychological states.

1 Introduction

As humans, we use narratives to express our representations of reality and make sense of the world (Bruner, 1990). Through written narratives, authors communicate events and reveal their unique ways of perceiving and interpreting reality. Personal narratives, the self-crafted stories authors create to make sense of their experiences, are rich sources for understanding how authors shape their identity and meaning-making (Labov and Waletzky, 1997).

Style has long been recognized as a key element in how authors express themselves through language. In everyday usage, style refers to a dis-

Table 1: Illustrative pipeline for our sequence-based framework. We first segment "I wake in a dark room. I feel a cold wind. I tell myself to move." into clauses, then identify features such as processes and participants for each clause. Each narrative is mapped to a symbolic sequence using an alphabet based on extracted features.

Clause	Process	Participants
I wake in a dark room	Action (**a**)	Actor
I feel a cold wind	Mental (**m**)	Senser, Phenomenon
I tell myself to move	Verbal (**v**)	Sayer, Recipient

Sequence: amv | **Substrings:** {am, mv}

tinctive manner of expression[1]. While this general definition encompasses all aspects of personal expression, our work narrows the focus to specifically examine how individuals linguistically encode their subjective experiences in personal narratives. Our approach complements existing works in stylometry (Neal et al., 2017) and stylistics (Wales, 2014) by providing a formal framework grounded in systemic functional linguistics (Halliday et al., 2014), which views language as a system of options for achieving communicative goals. This perspective aligns with how authors construct personal narratives, as they select from available linguistic resources to convey their subjective experience.

Scholarly investigations have long explored personal modes of reasoning and expression. For example, Husserl's phenomenology examined how intentional structures of consciousness shape lived experience (Husserl, 2012), Hadamard analyzed the cognitive processes underlying mathematical creativity (Hadamard, 1945), and Dilts modelled the problem-solving strategies of figures such as Leonardo da Vinci and Sherlock Holmes (Dilts, 1994). In his philosophy of style, Gilles-Gaston

---

[1]https://merriam-webster.com/dictionary/style

Granger defined style as the individual response to the difficulties faced by any structuring endeavor (Granger, 1968). While these qualitative frameworks richly describe personal modes of thought, they do not provide operational tools for capturing how such styles manifest in linguistic form.

A formalization of the intuitive notion of style in personal narratives is lacking. In therapeutic settings, where narrative reconstruction can facilitate healing (White and Epston, 1990), formalization could enable more precise identification of linguistic patterns associated with psychological states and support targeted interventions.

Tellier and Finkel (1995) propose a formal framework to characterize individual patterns between texts written by the same author. They propose a definition of linguistic style as lexical and syntactical patterns in the expressions of an intention. Drawing on these ideas, we develop a sequence-based framework to analyze how personal narratives convey subjective experience. Table 1 shows an illustrative pipeline for our framework. We represent narratives as sequences of linguistic choices, identifying recurring patterns in how authors encode their experiences. We do not claim to capture the full complexity of subjective experience in text. We aim to create a simple, accessible framework that researchers can use across disciplines and build upon in future studies. Our key contributions are:

1. A sequence-based framework defining style as patterns in sequences of linguistic choices grounded in systemic functional linguistics;

2. A methodology for automatically identifying patterns using sequence analysis;

3. A case study on dream narratives, showing how the analysis of patterns can reveal psychological insights and support therapeutic applications.

Our research task is to formalise style in personal narratives as patterns of linguistic choices that encode subjective experience, and to demonstrate that this formalisation yields interpretable empirical findings. Section 2 shows our categorization of linguistic features grounded in functional linguistics and how these features can be organized as sequences of symbols. Section 3 presents methods for analyzing patterns in these sequences. Section 4 demonstrates our framework through a case study on dream narratives. Finally, we conclude

and propose many applications (*e.g.*, authorship profiling, style-conditioned narrative generation) and theoretical extensions (applying methods from complexity science and formal language theory).

2 Categorizing Linguistic Features

What linguistic features are relevant for analyzing the communication of a subjective experience? This section presents our categorization of linguistic features based on functional linguistic theories. In these theories, language is seen primarily as a means to achieve communicative goals. Functional linguistics often draws on various disciplines, such as sociology, psychology, and cognitive science, to explore how language functions. We rely on a fundamental function of language: the communication of a subjective experience. It is through this notion that we define our concept of style. We introduce the various functional linguistic theories related to this function, then we present our categorization of features based on systemic functional linguistics.

Functions of language Karl Bühler's organon model represents one of the earliest attempts to categorize language functions (Bühler, 1990). He identified three fundamental functions: *representation* (encoding external reality), *expression* (revealing internal states), and *appeal* (influencing receivers of messages). The expressive function is particularly relevant to our analysis, as it specifically addresses how language manifests the internal state of the writer through linguistic choices, independent of both representational content and intended receiver effects. Roman Jakobson later expanded this model into a six-function framework (Jakobson, 2010).

Systemic functional linguistics Sharing conceptual roots with Bühler's and Jakobson's models of language, systemic functional linguistics, developed by Michael Halliday, offers a framework for analyzing how language functions as a meaning-making resource (Halliday et al., 2014). Rather than viewing language as a set of abstract rules, systemic functional linguistics examines it as a social semiotic system where meaning emerges through contextual linguistic choices. This approach is particularly valuable for analyzing personal narratives, as it provides tools for understanding how authors select and combine linguistic resources to construct meaning from their experiences. Systemic functional linguistics identifies three metafunctions: in-

Processes	Examples
Action: actions and events in the physical world.	[He]$_{Actor}$ [**takes**]$_{Action}$ [the valuable]$_{Affected}$ [Members of my cult]$_{Actor}$ [**have made**]$_{Action}$ [1500 euros]$_{Result}$ [I]$_{Actor}$ [**give**]$_{Action}$ [her]$_{Recipient}$ [a chance]$_{Range}$
Mental: internal experiences such as thoughts, perceptions, and feelings.	[We]$_{Senser}$ [**believe**]$_{Mental}$ [women are the leaders of change]$_{Phenomenon}$ [The moon]$_{Senser}$ [**sees**]$_{Mental}$ [the earth]$_{Phenomenon}$ [He]$_{Senser}$ [**disliked**]$_{Mental}$ [Gilbert's writing]$_{Phenomenon}$
Verbal: acts of communication.	[David]$_{Sayer}$ [**said**]$_{Verbal}$ ["the corrupt, criminals and money launderers"]$_{Verbiage}$
State: states of being, having, or existence.	There [**was**]$_{Existential}$ [a swimming pool]$_{Existent}$ [John]$_{Carrier}$ [**is**]$_{State}$ [an interesting teacher]$_{Attribute}$ [Hadrian's Wall]$_{Possessor}$ [**has**]$_{State}$ [something for everyone]$_{Possessed}$

Table 2: Processes with their participants.

terpersonal (how language is used to build and maintain social relationships), textual (how information is organized to create coherent messages), and ideational (how language represents experience through the transitivity system described below). Our framework particularly leverages the transitivity system, analyzing how authors encode their experiences through processes, participants, and circumstances.

Processes, participants, and circumstances The transitivity system, which shares similarities with Lucien Tesnière's concept of valency (Tesnière, 2015), provides a framework for analyzing how clauses represent experiences through the interaction of processes, participants, and circumstances. Processes (realized through verbal phrases), participants (realized through noun phrases), and circumstances (realized through adverbial groups or prepositional phrases) form the core of experiential representation. Inspired by Banks (2019), Table 2 presents our categorization of process types. Although there exists a finer categorization of processes (for example, the division of mental processes into cognitive, affective, and perceptive processes), we limit ourselves to a general categorization to save space. The definitions of participants and examples of circumstances can be found in Tables 3 and 4 (Appendix A).

We define *style* as patterns in choices within a system of linguistic features inspired by the transitivity system in systemic functional linguistics. We now need to explain what is a choice. To account for a subjective experience, an author makes a series of choices within a system of features.

3 Capturing Sequential Patterns

3.1 Formalizing Choice

Grammar as a system of meaningful choices Systemic functional linguistics views grammar as a network of meaningful choices (Halliday et al., 2014). This paradigmatic dimension emphasizes how linguistic units are selected from available options (de Saussure, 1916). When constructing clauses, authors select features from the transitivity system to encode their experience, making choices that fulfill specific communicative functions.

Formal representation of feature systems We represent each system of linguistic features as a finite set Σ containing available choices, also called an alphabet in formal language theory (Pin, 2021). For example, the alphabet of process types is $\Sigma_{process} = \{action, mental, verbal, state\}$. Using this alphabet, we can construct a word w with five letters, such as $w = action.mental.action.mental.state$ where . is the concatenation operator. Multiple alphabets can be combined through cartesian products, allowing the representation of simultaneous choices across different linguistic dimensions. For

example, we can have a finer-grained analysis with processes along with their tense and aspect: $\Sigma = \Sigma_{\text{process}} \times \Sigma_{\text{tense}} \times \Sigma_{\text{aspect}}$. Our case study on dream narratives is a simple proof of concept where we only focus on process types. Future works could incorporate more features, hence a larger alphabet, for a finer-grained analysis.

3.2 Sequence Analysis

Substrings and subsequences We analyze sequences to identify two types of recurring patterns: substrings and subsequences. A *substring* is any contiguous block of symbols occurring within a longer string. A *subsequence* is obtained by deleting zero or more symbols while preserving their relative order, so contiguity is not required. For example, $action.action$ is a subsequence of $action.mental.state.action$ but not a substring, whereas $action.mental$ is both a substring and a subsequence. Both notions come from the computer science terminology and underpin a large family of string-processing algorithms (Knuth et al., 1977). In formal language theory, substrings and subsequences are called *factors* and *subwords* (Berstel and Perrin, 2007).

Our study focuses on substrings, but our framework can be easily extended to consider subsequences. Identifying recurring substrings such as $action.mental$ within words can help detect patterns relevant to an author's style. For example, an author may tend to follow an action process with a mental process. By applying this framework to larger corpora, we can explore how authors structure their narratives, revealing the patterns that shape how subjective experience is communicated.

Analogy with computational biology As we aim to identify patterns in sequences of symbols, we can draw inspiration from concepts and methods developed in various fields that perform sequence analysis, such as computational biology (Waterman, 1995). For example, studying DNA or protein sequences in biology has led to the development of algorithms and similarity measures, notably through sequence alignment and complexity measures. These tools allow for identifying recurring patterns within sequences, quantifying the proximity between biological entities, and the inference of functional or evolutionary relationships. By analogy, if we replace nucleotides (A, C, G, T) with our linguistic features, we can employ these techniques to detect recurring patterns, measure the

stylistic similarity between different narratives, and highlight patterns specific to certain authors.

Similarity measures When comparing narratives, we focus on their stylistic patterns rather than their specific content by analyzing their associated sequences. While we could consider two narratives equal only if they have exactly the same sequence, this criterion is too strict for practical analysis. Instead, we need a more flexible way to measure how similar two sequences are. We propose using cosine similarity, which treats sequences as vectors in a high-dimensional space and measures the angle between them. For any two sequences s_1 and s_2, we first identify all possible substrings (contiguous subsequences) up to a certain length. We then count how frequently each substring appears in each sequence to create frequency vectors. The similarity is calculated as:

$$cos(s_1, s_2) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

where x_i and y_i represent the number of times a particular substring appears in s_1 and s_2 respectively. This measure ranges from 0 (completely different) to 1 (identical), with higher values indicating greater similarity. For example, consider two sequences of process types: $s_1 = amvma$ and $s_2 = amma$ (where $a = action$, $m = mental$, $v = verbal$). Their substrings of length one and two are: $s_1 : \{a : 2, m : 2, v : 1, am : 1, mv : 1, vm : 1, ma : 1\}$ and $s_2 : \{a : 2, m : 2, am : 1, mm : 1, ma : 1\}$. The cosine similarity between these sequences is 0.836. This high similarity reflects that they share many common substrings and have similar distributions of process types, even though s_1 contains a verbal process not present in s_2. Other similarity measures could be used, such as the Jaccard similarity or Euclidean distance.

Clustering We aim to find groups of narratives that reflect similar subjective experiences encoded in their patterns. We apply unsupervised clustering methods, particularly hierarchical agglomerative clustering (Ward, 1963), to automatically group similar sequences. This method constructs a hierarchy (or dendrogram) of clusters in a bottom-up manner. There is no need to specify the number of clusters as the methods produce a full hierarchy, allowing the analyst to choose a clustering level that best suits the needs. By examining the dendrogram, one can identify the number of clusters

that reflect meaningful stylistic distinctions. For instance, one might observe that certain narratives cluster together at a coarse level while finer-grained distinctions emerge deeper in the dendrogram.

Representative sequences One challenge in interpreting clustering results is to summarize each cluster in a way that highlights its most characteristic features. Representative sequences can summarize what stylistically unites the narratives in a cluster and offer a starting point for qualitative interpretation. To select representative sequences, we compute the sequence with the smallest average distance to all other sequences in the cluster.

In the next section, we apply sequence analysis methods to identify recurring patterns in dreams.

4 Case Study on Dream Narratives

To demonstrate the applicability of our framework, we perform a case study on dream narratives. By leveraging language models to automatically extract linguistic features at scale, we map each narrative into a sequence and uncover recurring patterns through sequence analysis. The patterns we identify offer insights into the psychological dimensions of how people articulate their experiences.

4.1 Data and Methodology

DreamBank corpus We apply our framework to dream narratives as they possess a narrative structure and represent intentional attempts to communicate subjective experience. Our analysis uses DreamBank, a corpus of thousands of dream narratives collected in the United States for scientific research. This English-language corpus has been employed in quantitative dream analysis (Domhoff and Schneider, 2008). The corpus is provided for academic research in dream and narrative analysis, and our usage strictly adheres to this intended purpose. From this corpus, we analyze five series of dreamers: *blind* (long-term blind dreamers, n=361), *ed* (a widower, n=139), *izzy* (a teenager, n=1091), *merri* (an artist, n=202), and *viet* (a Vietnam War veteran with post-traumatic stress disorder, n=566). A description of each series is accessible on the DreamBank website[2].

While this paper focuses on analyzing the *viet* series, we make the sequences of other series available. We construct a *norm* to establish a comparative baseline to compare how each series deviates

from a hypothetical average dreamer. This norm comprises a random sample of ten narratives from each series of DreamBank, resulting in a collection of 720 dream narratives.

Methodology Prior research has demonstrated the effectiveness of language models in analyzing dream narratives for character and emotion prediction (Bertolini et al., 2023; Cortal, 2024). Previous studies have also shown that systemic functional linguistics can be valuable for analyzing text corpora (Banks, 2002). Our work represents the first attempt to automate analysis from systemic functional linguistics using language models, focusing on the ideational function, particularly the transitivity system. We automatically extract processes, participants, and circumstances from dream narratives. This automation eliminates the costly and time-consuming manual extraction process that requires trained annotators and annotation campaigns.

Our methodology begins with dividing dream narratives into sentences using SpaCy segmentation model[3]. We then separate sentences into clauses using a language model with in-context examples (Dong et al., 2024). We do not differentiate between dependent and independent clauses. The next step involves extracting linguistic features using a language model with in-context examples from systemic functional linguistics books (Banks, 2019; Halliday et al., 2014). We use Llama 3.1 8B Instruct, a instruction tuned auto-regressive language model with 8-billion parameters (Grattafiori et al., 2024). This model was the best-performing 8B open-weights model at the time of our experiments[4]. We run our experiments using a Tesla V100 32GB for 80 hours.

As our sequence analysis depends on reliable extraction of linguistic features, we conducted a quantitative validation of the language-model annotator. We randomly sampled 50 clauses from the examples in Banks (2019) and Halliday et al. (2014). These clauses are considered gold-standard because they are fully analysed in the textbooks. Predicted processes, participants and circumstances were all correctly matched to the references.

We make our prompts and all extracted sequences available online for research use[5]. We acknowledge that our automatic extraction may

---

contain errors, such as misidentifying a mental process as a verbal process. Future work could evaluate our models against more expert annotations. Our framework benefits from the continuous improvement of language models. Llama 3.1 8B can be replaced with better language models in the future without altering the essence of our approach.

4.2 Analysis

We apply sequence analysis methods to identify recurring patterns in sequences of choices extracted from dream narratives. We aim to identify patterns characterizing how subjective experience is communicated in dream narratives.

To demonstrate the effectiveness of our framework and provide a simple case, we focus solely on the choice of processes. However, our approach can consider more complex configurations of linguistic features to identify finer-grained patterns. Our analysis focuses on the distribution of substrings and representative sequences of the war veteran with post-traumatic stress disorder. We discuss the psychological implications of our findings.

Distribution of substrings We compare the distribution of substrings between a series of dreamers and an average dreamer (*norm*). To save space, we only compare substrings of sizes one, two, and three. In Figure 1, we take *viet* as an example. Comparisons with other series are available in Appendix A. We also report other statistics such as per-series distribution of sequence lengths (Figure 8) and number of distinct substrings in the series (Figure 5). We compare the proportion of sequences containing a given substring between *viet* and *norm* using a Fisher exact test (with Holm–Bonferroni correction) (Fisher, 1935). We focus on the presence or absence of a substring within a sequence: each substring is treated as a categorical variable indicating whether or not it occurs in a sequence.

Odds ratio Figure 1 reports odds ratios with statistically significant results ($p < 0.05$). The odds ratio measures how much more (or less) likely it is to observe a given substring in one series compared to another. For example, Figure 1a shows that the presence of verbal processes is 40% more likely in the sequences of *viet* than in *norm*, as we have an odds ratio of 1.4. The presence of mental processes is 40% less likely in the sequences of *viet* than in *norm*, as we have an odds ratio of 0.6. Figures 1a, 1b, and 1c show a preference for *viet* to remain in a verbal process, as indicated by substrings such as
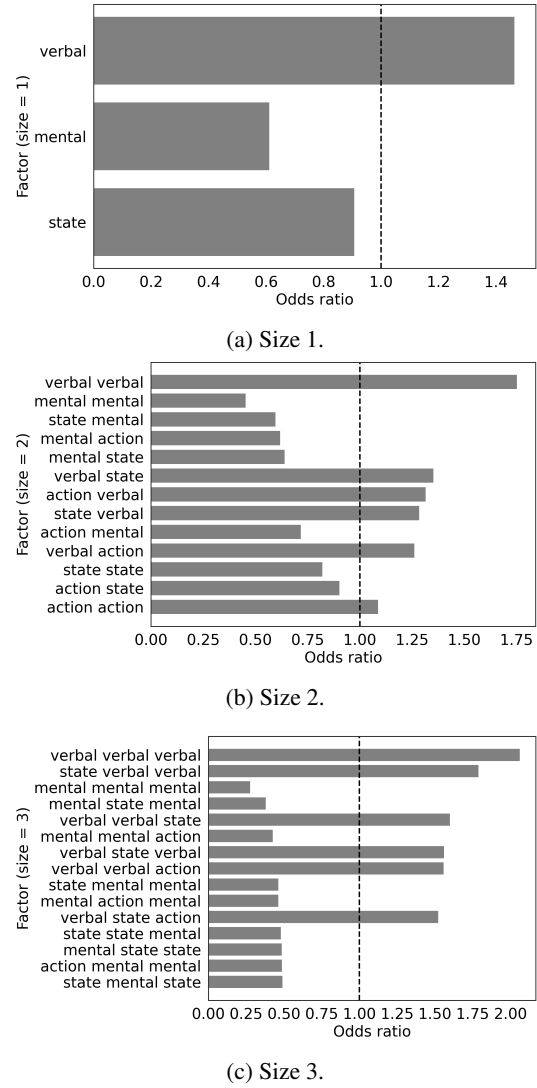

(a) Size 1.

(b) Size 2.

(c) Size 3.

Figure 1: Top substring between *viet* and *norm*.

verbal.verbal and *verbal.verbal.verbal* with high odds ratios (respectively 2.00 and 1.75). We save space by showing only substring sizes up to three. However, we analyze other sizes that reveal interesting patterns. For example, some substrings do not appear in *viet* but appear in *norm*, such as *mental.state.mental.mental*.

Representative sequences We perform hierarchical agglomerative clustering on the sequences of *viet* to find representative sequences that cover an important portion (80%) of all sequences while minimizing intra-cluster distances. We use Ward linkage and cosine similarity. The silhouette score, which measures how similar sequences are to their own cluster compared to other clusters, reaches its maximum value with substrings of lengths up to three and when partitioning the dendrogram into two clusters (shown in orange and green in
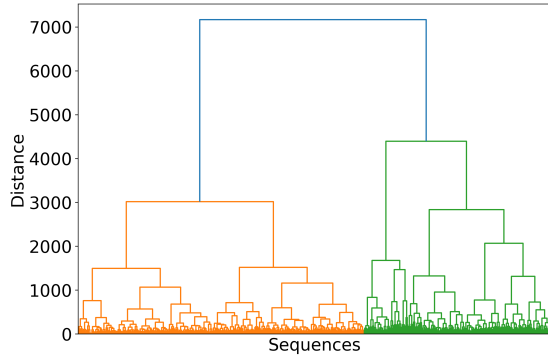
Figure 2: Dendrogram for *viet* based on substrings of size one, two, and three.

Figure 2). This suggests two distinct patterns in how the veteran structures his subjective experience in dream narratives. We identified a representative sequence for each cluster by calculating the sequence with the minimal average distance to all other sequences within its cluster. The first representative sequence (*savamasasaaa-maaasavvvaaaaaaavssaaaaa*) demonstrates a dominance of action processes (23 occurrences) with minimal mental processes (2) and covers 274 sequences with an average distance of 8 to other sequences in its cluster. The second representative sequence (*ssssavaavssvsavvvvsmasasaasasaa-maamvmsss*) reveals a more balanced pattern, featuring both action (13 occurrences) and state processes (16 occurrences) while maintaining relatively few mental processes (4), and covers 179 sequences with an average distance of 21. These patterns suggest that the veteran follows two templates: a highly action-oriented structure or a more varied structure alternating between state and action processes.

Psychological interpretations We can bridge our analysis with therapeutic goals. Identifying low- or high-odd substrings could serve as diagnostic markers or guide therapy interventions. As therapists and patients collaborate to restructure narratives, certain stylistic elements can help target specific patterns for modification. For instance, mental processes occurring more frequently than average in an author's narratives might indicate an intensified inward focus. Conversely, a relative scarcity of such patterns could suggest a more action-oriented or externally focused stance. In our example, *viet* primarily frames his experiences through action and verbal rather than mental processes. This may align with studies about how combat trauma can affect cognitive and emotional processing (American Psychiatric Association, 2013), though further research would be needed to confirm this interpretation. We aim to show that our formalization of style can help authors learn to reframe their experiences, potentially facilitating trauma processing.

Our analysis relies solely on process types to show a simple proof of concept. Future works will include a fine-grained sequence analysis considering processes, participants, and circumstances. These building blocks can be further specified with features with a functional role, such as the duration and completeness of processes (realized by tense and aspect) and the concreteness of participants.

5 Related Work

Systemic Functional Linguistics parsing. Early efforts to operationalise Halliday's transitivity system were almost exclusively rule-driven (Honnibal and Curran, 2007; O'Donnell, 2019). They map phrase-structure parses onto SFL networks, but coverage is brittle outside the newswire sentences on which the rules were crafted. Recent graph-based pipelines transform dependency trees into reduced mood and transitivity networks (Costetchi, 2013). Supervised approaches improve verb-class recognition provided that expert-annotated data exist (Yan and Webster, 2014). A recent survey concludes that SFL parsers remain "experimental, domain-sensitive, and labour-intensive to adapt" (O'Donnell, 2019).

Our framework departs from this rule-engineering paradigm. We prompt a large language model to jointly label processes, participants, and circumstances. This few-shot approach removes the need for handcrafted grammars or verb lexicons, and it scales to thousands of narratives without additional annotation effort.

Computational analysis of dream narratives. Quantitative dream research traditionally depends on manual coding systems, which is labour-intensive and difficult to scale. Rule- and lexicon-based methods map dream words to hand-curated dictionaries or WordNet synsets to identify characters, interactions, or emotions (Bulkeley and Graves, 2018; Fogli et al., 2020; Mallett et al., 2021; Zheng and Schweickert, 2021). Distributional approaches embed dreams in a semantic space and cluster them by themes such as flight or fight (Altszyler et al., 2017; Sanz et al., 2018; Gutman Music et al., 2022). Hybrid systems add classifiers on top

7317

of lexicon scores to predict overall sentiment or social themes (McNamara et al., 2019; Yu, 2022). Language-model approaches are only emerging: Bertolini et al. (2023) fine-tune BERT to detect the presence of emotions, and Cortal (2024) fine-tune a sequence-to-sequence transformer to predict character and emotion. All of these approaches mainly focus on *what* is said (*e.g.*, entities, emotions, or thematic content). In contrast, we target *how* subjective experience is constructed linguistically.

6 Conclusion and Perspectives

We present a novel framework that formalizes style in personal narratives as patterns in linguistic choices used to communicate subjective experience. By combining functional linguistics, sequence analysis methods, and psychological interpretations of sequential patterns, we analyze how authors encode their experiences through language. Our framework demonstrates the potential for uncovering psychologically meaningful patterns in dream narratives. The analysis of a war veteran reveals distinctive patterns in process types, particularly a prevalence of verbal processes over mental ones, suggesting links between linguistic choices and trauma processing. Our work opens directions for future research.

Authorship profiling Authorship profiling aims to infer an author's demographic and psychological characteristics based on their written texts. Applications range from forensic linguistics (Gibbons, 2003) and literary attribution (Ledger and Merriam, 1994) to online security (Argamon et al., 2009) and bot detection (Ferrara et al., 2016). Our framework could enhance existing approaches to authorship attribution, particularly for texts where patterns related to subjective experience diagnose authorship. We can identify signature patterns (*e.g.*, distinctive substrings) that characterize an author's unique way of constructing narratives.

Style-conditioned narrative generation Language models have demonstrated increasing utility in psychotherapeutic applications (Cortal et al., 2023; Bonard and Cortal, 2024), particularly in text transformation tasks such as cognitive reframing of negative thoughts (Sharma et al., 2023) and generating positive perspectives (Ziems et al., 2022). While most research in controllable text generation has focused on high-level attributes like formality and sentiment (Jin et al., 2022; Troiano et al., 2023),

recent work has shifted toward controlling fine-grained linguistic attributes (Alhafni et al., 2024). Our sequence-based framework offers a novel approach to controlling fine-grained attributes related to communicating subjective experience. While our current work focuses on structuring linguistic choices to characterize style from narratives, we could explore the inverse process: generating narratives from sequences of choices. This inverse mapping could enable style-conditioned narrative generation, offering new possibilities for therapeutic interventions or literary writing.

Applying methods from complexity science and formal language theory While our current analysis focuses on substrings, analyzing subsequences might offer more flexible pattern detection as it involves non-contiguous patterns. Moreover, we could draw from complexity measures to quantify redundancies in sequences. For example, the Lempel-Ziv measure evaluates how efficiently a sequence can be parsed into non-redundant substrings (Lempel and Ziv, 1976), a low complexity indicating predictable patterns. These measures can reveal psychological insights: low complexity may indicate fixed mental representations or recurring cognitive schemas, particularly relevant in trauma cases, while higher complexity might suggest greater cognitive flexibility. This aligns with research applying complexity science to understand mental disorders (Carhart-Harris, 2018; Hipólito et al., 2023; Hong et al., 2015).

Limitations

Our approach relies on language models to segment and annotate clauses. Although automation enables large-scale analysis, it can introduce inaccuracies when models misclassify processes or participants. Such inaccuracies can undermine the reliability of observed patterns. Improvements in model performance, as well as future efforts to validate or correct automated annotations using expert-annotated data, would help ensure more trustworthy results.

While we connect patterns in linguistic choices to psychological interpretations, these connections remain correlational and descriptive. Identifying a higher or lower prevalence of certain processes (*e.g.*, mental or verbal) may hint at underlying psychological states, but these inferences are speculative. Validation against clinical assessments would be necessary to strengthen claims about the therapeutic utility of our stylistic analyses.

Despite these limitations, our approach lays the groundwork for a sequence-based, linguistically informed analysis of personal narratives. Future studies can address these limitations by expanding the range of linguistic features, refining extraction methods, and validating psychological interpretations with clinical data.

Ethics Statement

Our study involves autobiographical dream narratives that may reveal mental-health information about the dreamers. All texts are obtained from the publicly available DreamBank database and are anonymized. All dreamers are identified only by pseudonyms supplied by the original collectors. The dreamers have given their consent to appear in the database. No attempt was made to infer individual clinical diagnoses beyond the descriptive case study, and we emphasise that stylistic patterns should never be treated as diagnostic in isolation.

The automatic annotations are produced with an open-weights language model (Llama 3.1 8B). Because such models can encode socially biased associations, we release the prompts, hyperparameters and extracted sequences to enable third-party auditing[6]. Potential downstream uses (*e.g.*, automated triage) might amplify misclassification harms. We stress that our framework is intended for exploratory research or therapist-in-the-loop settings, not fully autonomous clinical deployment.

Acknowledgments

References

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized Text Generation with Fine-Grained Linguistic Control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition*, 56:178–187.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123.

David Banks. 2002. Systemic Functional Linguistics as a model for text analysis. *ASp. la revue du GERAS*, (35-36):23–34. Number: 35-36 Publisher: Groupe d'étude et de recherche en anglais de spécialité.

David Banks. 2019. *A Systemic Functional Grammar of English: A Simple Introduction*. Routledge.

Jean Berstel and Dominique Perrin. 2007. The origins of combinatorics on words. *European Journal of Combinatorics*, 28(3):996–1022.

Lorenzo Bertolini, Valentina Elce, Adriana Michalak, Giulio Bernardi, and Julie Weeds. 2023. Automatic scoring of dream reports' emotional content with large language models. *Preprint*, arXiv:2302.14828.

Constant Bonard and Gustave Cortal. 2024. Improving language models for emotion analysis: Insights from cognitive science. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 264–277, Bangkok, Thailand. Association for Computational Linguistics.

Jerome Bruner. 1990. *Acts of meaning*. Acts of meaning. Harvard University Press, Cambridge, MA, US.

Kelly Bulkeley and Mark Graves. 2018. Using the LIWC program to study dreams. *Dreaming*, 28(1):43–58.

Karl Bühler. 1990. *Theory of Language: The Representational Function of Language*. John Benjamins.

Robin L. Carhart-Harris. 2018. The entropic brain - revisited. *Neuropharmacology*, 142:167–178.

Gustave Cortal. 2024. Sequence-to-sequence language models for character and emotion detection in dream narratives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14717–14728, Torino, Italia. ELRA and ICCL.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

---

[6] https://github.com/gustavecortal/formalizing-style-in-personal-narratives

Eugeniu Costetchi. 2013. A method to generate simplified systemic functional parses from dependency parses. In *Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing 2013)*, pages 68–77, Prague, Czech Republic. Charles University in Prague, Matfyzpress.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.

Robert Dilts. 1994. *Strategies of Genius, Volume 1: Aristotle, Sherlock Holmes, Walt Disney, Wolfgang Amadeus Mozart*. Meta Publications, Capitola, CA.

G. William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition*, 17(4):1238–1247.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. 2016. Detection of Promoted Social Media Campaigns. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):563–566.

R. A. Fisher. 1935. *The design of experiments*. The design of experiments. Oliver & Boyd, Oxford, England.

Alessandro Fogli, Luca Maria Aiello, and Daniele Quercia. 2020. Our dreams, our selves: automatic analysis of dream reports. *Royal Society Open Science*, 7(8):192080.

John Gibbons. 2003. *Forensic Linguistics: An Introduction to Language in the Justice System*. Wiley.

Gilles-Gaston Granger. 1968. *Essai d'une philosophie du style*. A. Colin, Paris.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*.

Maja Gutman Music, Pavan Holur, and Kelly Bulkeley. 2022. Mapping dreams in a computational space: A phrase-level model for analyzing fight/flight and other typical situations in dream reports. *Consciousness and Cognition*, 106:103428.

Jacques Hadamard. 1945. *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press, Princeton, N.J.

M.A.K. Halliday, Christian M.I.M. Matthiessen, Michael Halliday, and Christian Matthiessen. 2014. *An Introduction to Functional Grammar*. Routledge.

Inês Hipólito, Jonas Mago, Fernando E Rosas, and Robin Carhart-Harris. 2023. Pattern breaking: a complex systems approach to psychedelic medicine. *Neuroscience of Consciousness*, 2023(1):niad017.

Kai Hong, Ani Nenkova, Mary E. March, Amber P. Parker, Ragini Verma, and Christian G. Kohler. 2015. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry Research*, 225(1):40–49.

Matthew Honnibal and James R. Curran. 2007. Creating a systemic functional grammar corpus from the Penn treebank. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pages 89–96, Prague, Czech Republic. Association for Computational Linguistics.

Edmund Husserl. 2012. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: First Book: General Introduction to a Pure Phenomenology*, illustrated edition. Springer Science & Business Media, Dordrecht.

Roman Jakobson. 2010. Linguistics and Poetics. In *Volume III Poetry of Grammar and Grammar of Poetry*, pages 18–51. De Gruyter Mouton.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Donald E. Knuth, James H. Morris, Jr., and Vaughan R. Pratt. 1977. Fast Pattern Matching in Strings. *SIAM Journal on Computing*, 6(2):323–350. Publisher: Society for Industrial and Applied Mathematics.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History*, 7(1-4):3–38. Place: Netherlands Publisher: John Benjamins.

Gerard Ledger and Thomas Merriam. 1994. Shakespeare, Fletcher, and the Two Noble Kinsmen. *Literary and Linguistic Computing*, 9(3):235–248.

A. Lempel and J. Ziv. 1976. On the Complexity of Finite Sequences. *IEEE Transactions on Information Theory*, 22(1):75–81.

Remington Mallett, Claudia Picard-Deland, Wilfred Pigeon, Madeline Wary, Alam Grewal, Mark Blagrove, and Michelle Carr. 2021. The relationship between dreams and subsequent morning mood using self-reports and text analysis. *Affective Science*, 3(2):400–405.

Patrick McNamara, Kelly Duffy-Deno, Tom Marsh, and Thomas Jr. Marsh. 2019. Dream content analysis using artificial intelligence. *International Journal of Dream Research*, 12(1):42–52.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.*, 50(6):86:1–86:36.

Mick O'Donnell. 2019. Continuing issues in SFL. In Geoff Thompson, Wendy L. Bowcher, Lise Fontaine, and David Schönthal, editors, *The Cambridge Handbook of Systemic Functional Linguistics*. Cambridge University Press, Cambridge. Online first 2016.

Jean-Eric Pin, editor. 2021. *Handbook of Automata Theory*, volume 2. EMS Press.

Camila Sanz, Federico Zamberlan, Earth Erowid, Fire Erowid, and Enzo Tagliazucchi. 2018. The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Frontiers in Neuroscience*, 12.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Isabelle Tellier and Alain Finkel. 1995. Cognitive style of decision making narrations. In Heide Endres-Niggemeyer, editor, *The Cognitive Level*, Duisburg LAUD Series of Cognitive Linguistics, pages 41–59.

Lucien Tesnière. 2015. *Elements of Structural Syntax*. John Benjamins Publishing Company, Amsterdam.

Enrica Troiano, Aswathy Velutharambath, and Roman Klinger. 2023. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, 29(4):849–908.

Katie Wales. 2014. *A Dictionary of Stylistics*. Routledge.

Joe H. Ward. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Michael S. Waterman. 1995. *Introduction to Computational Biology: Maps, Sequences and Genomes*. CRC Press.

Michael White and David Epston. 1990. *Narrative Means To Therapeutic Ends*. WW Norton.

Hengbin Yan and Jonathan J. Webster. 2014. Automatic labelling of transitivity functional roles. *Journal of World Languages*, 1(2):157–170.

Calvin Kai-Ching Yu. 2022. Automated analysis of dream sentiment—the royal road to dream dynamics? *Dreaming*, 32(1):33–51.

Xiaofang Zheng and Richard Schweickert. 2021. Comparing hall van de castle coding and linguistic inquiry and word count using canonical correlation analysis. *Dreaming*, 31(3):207–224.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing Positive Perspectives with Text Reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

In the next page, we describe processes and participants. We also provide more statistical analysis of sequences such as the number of distinct substrings in the series, top substrings for each series, and the distribution of sequence lengths.
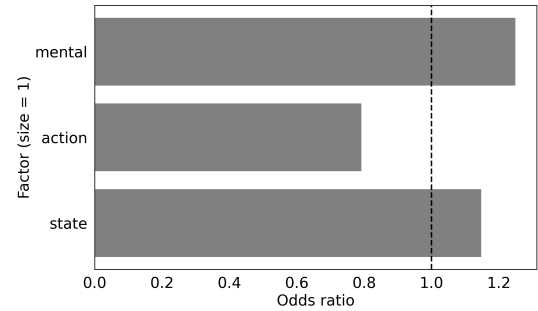
Processes	Participants
Action: actions and events in the physical world.	*Actor*: performs the action *Affected*: receives the impact of the action *Result*: outcome or creation from the process *Recipient*: entity to whom the action is directed
Mental: internal experiences such as thoughts, perceptions, and feelings.	*Senser*: experiencer of the process *Phenomenon*: what is being experienced
Verbal: acts of communication.	*Sayer*: communicator of the message *Verbiage*: content of the message *Recipient*: receiver of the message
States: states of existence, being or having.	*Existent*: entity that exists *Carrier*: entity linked to an *Attribute* *Attribute*: quality or descriptor of the *Carrier* *Possessor*: owner of something *Possessed*: item owned by the *Possessor*

Table 3: Processes and their participants.

Circumstances	Examples
Time: When?	She arrived **yesterday**. He visits his grandmother **every week**. They worked **for three hours**.
Place: Where?	The meeting will be held **in the conference room**. They walked **to the park**. He came **from New York**.
Manner: How?	She spoke **softly**. He fixed the car **with a wrench**. He loves her **very much**.
Cause: Why?	She stayed home **because of the rai**n. He went to the store **to buy milk**.
Accompaniment: With whom?	She went to the movie **with her friends**. He completed the project **without any help**.
Matter: About what?	They had a discussion **about the new policy**. She wrote a report **on climate change**.
Role: As what?	She attended the meeting **as a representative**. He acted **as the spokesperson**.

Table 4: Types of circumstances with examples.

Series	Size	Count	Sig. Diff.
viet	1	4	3
	2	16	13
	3	64	35
	4	253	59
	5	864	73
	6	2017	68
	7	3337	42
	8	4283	23
	9	4773	13
blind	1	4	3
	2	16	5
	3	64	12
	4	253	27
	5	786	55
	6	1570	47
	7	2137	33
	8	2285	8
	9	2202	6
izzy	1	4	3
	2	16	12
	3	64	39
	4	255	79
	5	930	101
	6	2520	98
	7	4828	92
	8	7007	58
	9	8568	26
ed	1	4	4
	2	16	8
	3	64	26
	4	254	47
	5	783	73
	6	1406	93
	7	1722	67
	8	1790	39
	9	1727	21
merri	1	4	2
	2	16	9
	3	64	21
	4	246	23
	5	740	22
	6	1564	40
	7	2447	41
	8	3157	23
	9	3610	17

Table 5: **Series** is the series name, **Size** is the substring size, **Count** is the number of distinct substrings in the series, **Sig. Diff.** is the count of substrings with significant differences relative to the norm.

(a) Size 1.
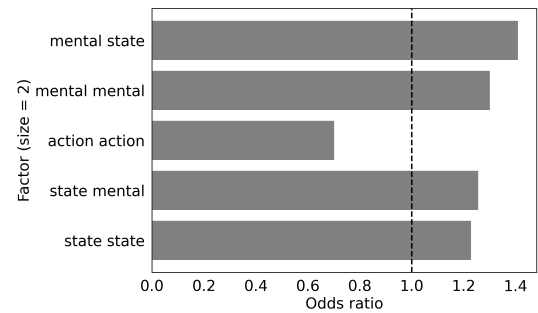
(b) Size 2.
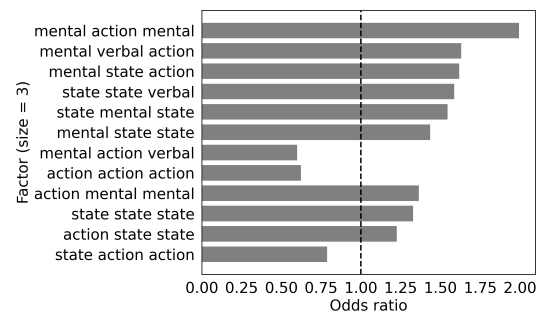
(c) Size 3.
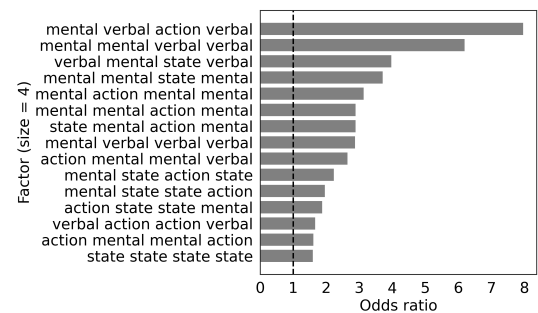
(d) Size 4.
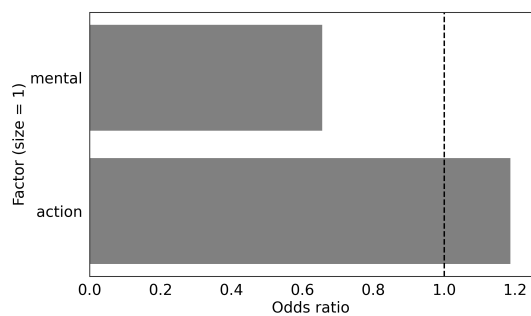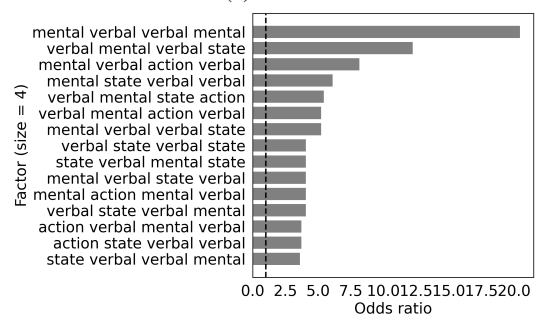
Figure 3: Top substring between *blind* and *norm*.
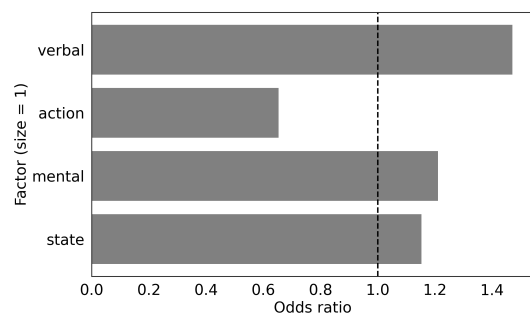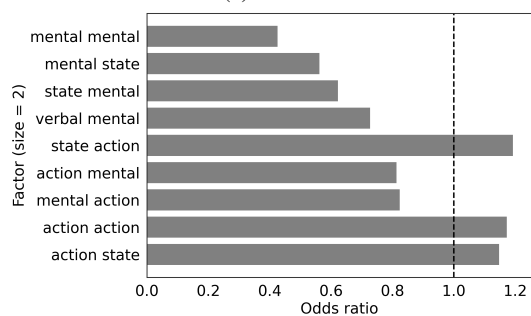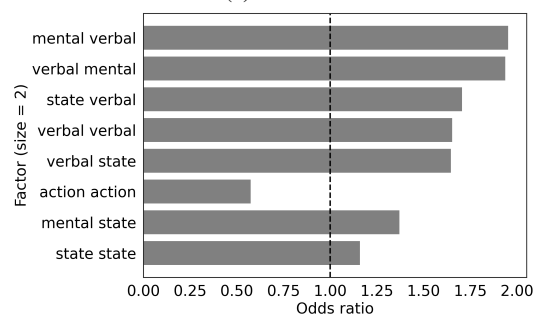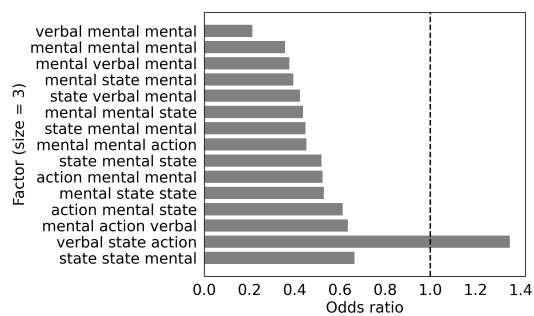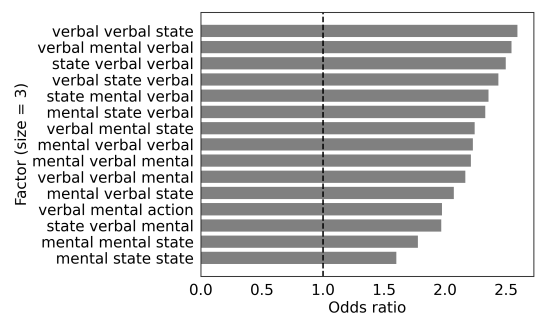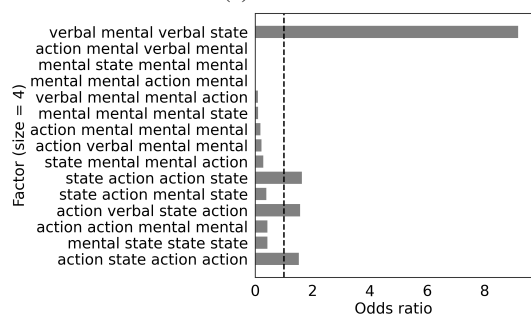
(a) Size 1.

(b) Size 2.

(c) Size 3.

(d) Size 4.

Figure 4: Top substring between *merri* and *norm*.

(a) Size 1.

(b) Size 2.

(c) Size 3.

(d) Size 4.

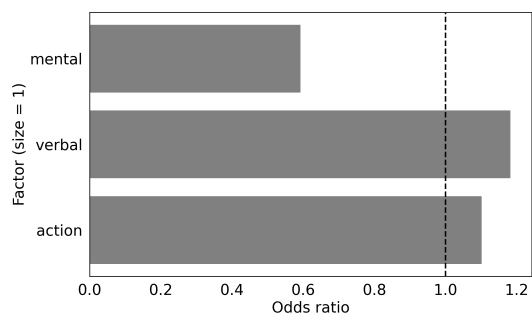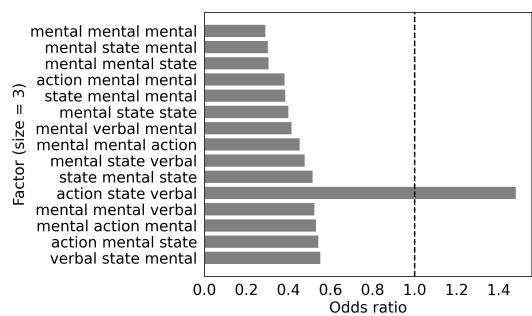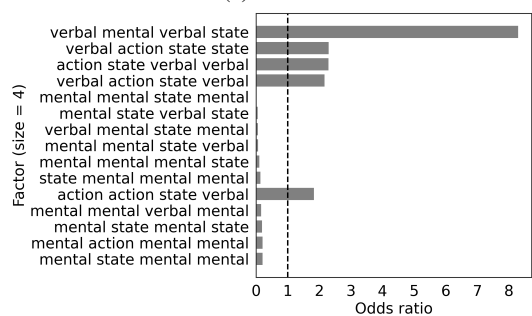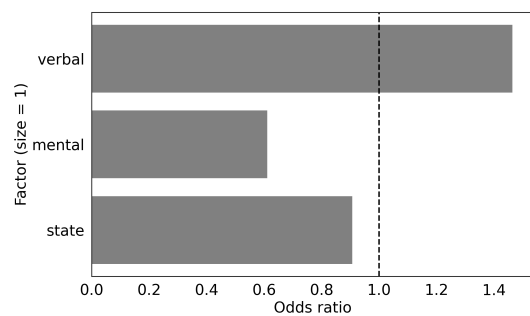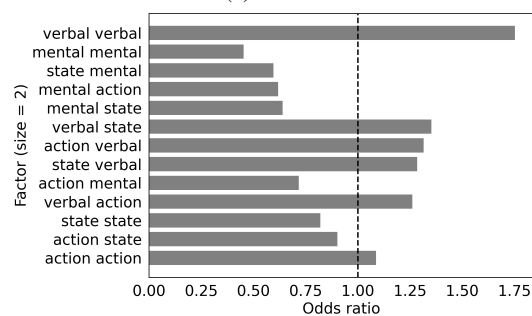Figure 5: Top substring odds ratio between *ed* and *norm*.
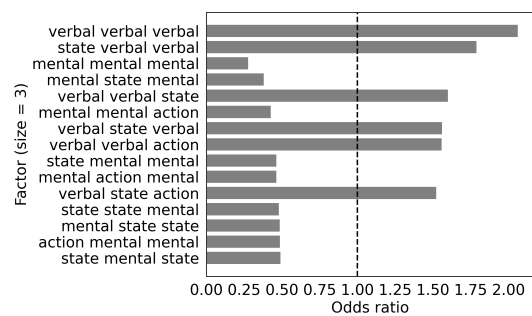
(a) Size 1.
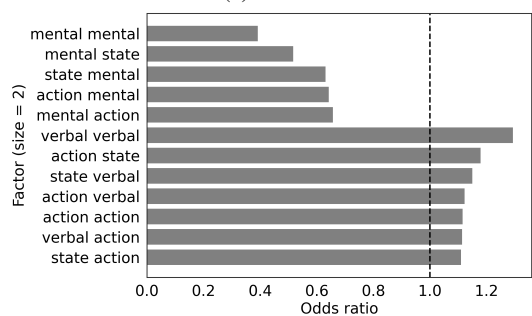
(b) Size 2.

(c) Size 3.

(d) Size 4.

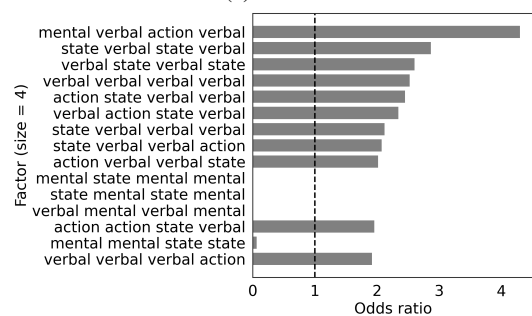Figure 6: Top substring between *izzy* and *norm*.
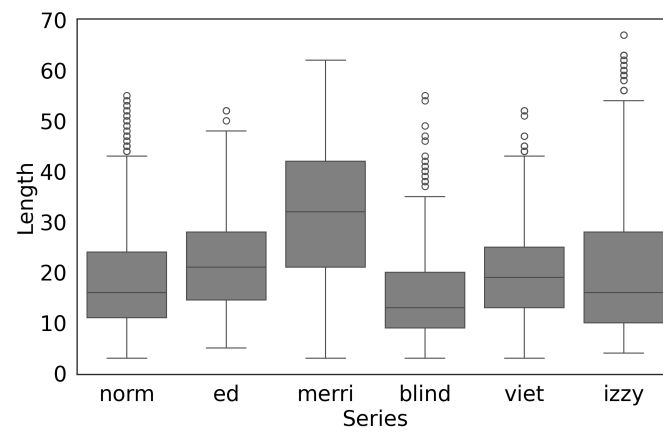
Figure 7: Top substring between *viet* and *norm*.

Figure 8: Per-series distribution of sequence lengths.