Metanarratives in the Fashion Domain

submitted in partial fulfillment for the degree of master of science

Zoë Oggel
13383337

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 24.07.2026

	UvA Supervisor	**UvA Supervisor**
Title, Name	Alexandra Barancová	Giovanni Sileno
Affiliation	University of Amsterdam	University of Amsterdam
Email	a.barancova@uva.nl	g.sileno@uva.nl



Abstract

This thesis applies computational narrative analysis to fashion criticism, asking how metanarratives can be detected in runway reviews. The corpus consists of 6,501 runway reviews spanning from 1999-2014. Two complementary layers are modeled. First, a RollingLDA topic model tracks thematic content over time, finding 10 topics which are then annotated over the whole corpus by an LLM based on a sample coding study. Second, per-review event sequences are extracted via dependency parsing and a clustering 30-type verb alphabet, nested into coarser rhetorical moves (4) and agency (6) schemes. Inferential tests over label-permutation and shuffle nulls isolate order, composition, and group signal. The results show that thematic composition is strongly designer-driven, whereas event structure carries little compositional signal which can be explained by the author of the review. Order adds nothing beyond composition. Fashion reviews thus emerge as a genre where descriptive themes belong to designers and the described action belongs to the reviews.

Keywords

computational narrative analysis, metanarratives, fashion reviews, topic modeling, event-based modeling, cultural analytics, NLP

Github Repository

https://github.com/Zoe-O16/Thesis.git

1 Introduction

Fashion is a complex system consisting of interwoven levels of granularity, covering daily personal choices considering practicality and expression, as well as global infrastructures and cultural indicators. The act of getting dressed every day implicitly contains an array of cultural, ethical, environmental, political, and economic questions [32]. Runway shows have traditionally been responsible for the constant innovation in fashion, both reflecting societal values and influencing them. As clothing is a near-universal aspect of everyday life, the ways in which it is designed, presented, and interpreted carry cultural significance beyond aesthetics alone.

Runway shows are ephemeral events, which leads to their cultural meaning being sustained through textual and visual mediation in order for it to reach far beyond the often exclusive spaces they take place in. Runway reviews do not merely describe garments, but place collections within the wider context of fashion discourse. In doing so, they help shape how fashion is interpreted and positioned over time, making them an important lens through which broader social and cultural developments can be studied [40].

Considering the creation of cultural meaning through text, the mechanism of narrative comes to mind. Narrative has been argued to be one of the primary cognitive and cultural modes through which people interpret reality and create shared meanings [14]. Beyond simple reporting of events, narratives provide frameworks that shape how societies interpret and negotiate affective, cultural, and ideological values [23].

Within cultural discourse, individual narratives are often structured by broader shared story frameworks. Research on meta- or master narratives describes these as culturally shared story templates that shape how smaller narratives are constructed and interpreted [53]. Metanarratives are therefore not located in any single text, but emerge across many individual narratives: recurring patterns become visible only when narratives are read in aggregate rather than in isolation [28]. Applied to fashion discourse, runway reviews may therefore reflect recurring higher-level narrative patterns that frame collections within broader cultural storylines.

The field of critical fashion studies aims to make sense of the evolving landscape of fashion from a cultural studies perspective [32]. However, fashion narratives are produced at a scale that exceeds manual analysis. Per season, hundreds of designers show their work across multiple cities. While qualitative analysis provides depth, it is less suitable for studying long-term, large-scale trends. However, a computational narrative analysis approach would be well suited for the task, applying NLP techniques to identify patterns at scale.

Despite the narrative nature of fashion discourse, computational analyses of fashion texts have primarily focused on trend detection [3, 8]. At the same time, computational narrative research has developed methods for identifying narrative structures within large corpora, including event-based narrative modeling and thematic analysis [6, 16]. These methods, however, have largely been developed to characterize narratives within individual texts, and offer little guidance on how to aggregate across a corpus to identify the higher-level, culturally shared patterns that constitute metanarratives.

This reveals both a conceptual and methodological gap. Conceptually, while fashion is widely understood as a cultural communication system, it remains unclear how narratives manifest in runway reviews and what constitutes a higher-level narrative within this domain. Methodologically, existing computational fashion research has focused largely on trends and sentiments, leaving open how narrative structures can be operationalized and analyzed quantitatively across large collections of fashion texts across time. Furthermore, computational narrative methods have largely operated at the level of individual texts, offering little guidance on how to aggregate narratives across a corpus to detect the higher-level, culturally shared patterns that constitute metanarratives.

This study aims to address these gaps by analyzing a dataset of runway reviews, spanning from 1999 to 2014, as a longitudinal narrative corpus. Using computational narrative analysis methods, it investigates whether recurring metanarrative patterns can be identified across designers over time, leading to the following central research question:

How can metanarratives in fashion runway reviews be identified computationally?

- 1. How can computational narrative analysis methods be used to detect metanarratives within fashion discourse?
- 2. What recurring narrative structures emerge across runway reviews?

- 3. How do these structures evolve across designers and years?

2 Related Work

2.1 Metanarratives

In narratology, a narrative can be understood as a sequence of connected events organized into a meaningful structure that conveys interpretation and context [38]. To describe overarching stories that smaller narratives can be generalized to, several strands of research use similar concepts, such as master narratives or metanarratives. Research in discourse and communication studies describes master narratives as larger cultural stories that shape how individual narratives are constructed and understood. For example, Smith and Dougherty [53] define master narratives as broader cultural narratives that silently shape local narratives by imposing cultural norms and expectations. In this view, narratives are not isolated accounts but are structured by widely shared cultural templates. Similarly, research on identity narratives shows that these dominant cultural stories function as collective frameworks through which individuals interpret their identities and experiences. Individuals often position themselves either in alignment with or in opposition to such master narratives, demonstrating their role as cultural reference points that guide interpretation and behavior [22]. The paper by Kaplan et al. [28] seems to be a singular example of operationalizing and studying metanarratives quantitatively. They define metanarratives as structures made of three types of components: actions, actors, and values. The authors argue that while specific narratives are unique, they can be abstracted up to general components that can be compared. In order to study this, the authors hand-code constitution preambles and infer the metanarrative via the most salient type of each component.

Taken together, this leads to a definition of metanarratives as recurring higher-level narrative patterns that organize texts within broader cultural storylines, which are composed of individual narrative components.

2.2 Fashion and Narratives

The narrative components (actions, actors, values) included in the study by Kaplan et al. [28] make sense for a political paper, but might not fully translate to the fashion domain. Despite there being no formal definition of a metanarrative in the fashion domain, fashion has been argued to be a tool of communication, with its meaning embedded in a broader cultural context[9, 15]. This has roots in fashion theory. Barthes' *Fashion System* discussed how fashion translates into language and functions as a system of signs, assigning fashion magazines a central role in this process [10, 35]. Therefore, it can be expected that a runway review not only communicates something about what appeared on the runway, but also how the meaning of the collection extends beyond its environment.

What a review foregrounds is a description rather than explicit value: a review is above all an account of what a collection looked like, and the evolution of those descriptions over time is itself the main signal in fashion discourse. Therefore, a descriptive theme would be a better fit than value, defining metanarratives in fashion as higher-level narrative patterns that organize texts within broader cultural storylines, composed of actors, actions, and descriptive themes.

2.3 Narrative Analysis Methods

To analyze metanarratives at scale within the fashion domain, the individual components as well as the higher-level structure of a metanarrative require a definition. The field of computational narrative understanding offers a diverse range of methods for capturing both individual narratives and higher-level patterns.

Piper [41] organizes computational narrative understanding across four levels: elements, modes, shapes, and collective schemas. The level of elements distinguishes diegetic elements of the story (agents, actions, objects, setting, time) from heterodiegetic elements of the telling (teller, recipient, mode, social situation). Modes concern the medium of the narrative (text, image, sound), and shapes concern the temporal patterning of elements as pathways, arcs, or other forms. Collective schemas concern the recurring story types (decline, progress, redemption) into which individual stories aggregate.

2.3.1 Narrative Extraction. Computational work has largely addressed the element level of narratives, through tasks such as character, event, and relation extraction [41]. A survey by Santana et al. [49] describes narrative extraction as a process that involves identifying narrative elements (events, participants, space, and time), linking these elements, and representing the resulting narratives. The extraction of events would fit the *actions* component of the metanarrative, and the extraction of participants would fit the *actors* component.

2.3.2 Event-Based Modeling. Once events have been extracted, a narrative can be characterized in two ways: the order in which events occur, and the composition of the event set.

Event Order. Existing work on computational narrative analysis focuses on ordering. Chambers and Jurafsky [16] introduce narrative event chains, an unsupervised method that learns coherent event sequences by tracking verbs and their coreferring arguments across documents. Their approach demonstrates that narratives can be modeled as recurring structures consisting of events and participant roles, allowing systems to detect patterns in how actions unfold over time without requiring predefined event types.

Event Composition. Because reviews are often no more than a few hundred words, what is mentioned in them might matter more than in what order it is mentioned. However, composition has received less attention in computational narrative analysis. Kaplan et al. [28] come closest, coding the relative share of actions across their dataset. However, the coding is done manually, and the proportions are collapsed to a single most-salient type per component.

2.3.3 Topic Modeling. While event-based approaches capture narrative progression within individual texts, topic modeling enables the identification of recurring thematic structures across large corpora. Topic models such as Latent Dirichlet Allocation (LDA) [12] detect clusters of co-occurring terms that represent latent topics within a dataset. These methods have been widely used to analyze large collections of textual data and identify dominant themes in online discourse and review datasets [13, 31].

When discourse evolves over time, static topic models are insufficient, as they assume that thematic structures remain constant across the full corpus. Two approaches address this limitation. Dynamic Topic models [11] extend LDA by modeling topics as evolving via a Gaussian random walk between consecutive time slices. Because inference is performed jointly across the full time series, however, later documents can influence the representation of earlier periods. RollingLDA, another approach for longitudinal topic modeling, addresses this by fitting an initial model on an early segment of the corpus and updating it sequentially for each following time slice, using a rolling memory window of preceding documents as an anchor. This allows for more tolerance for abrupt vocabulary change [44]. Topic assignments are fixed to prevent information leakage from future documents [46], which also allows for more reliable evaluation over the corpus.

Furthermore, recent work has begun using large language models as annotation tools for narrative analysis. For example, studies of online survivor narratives employ GPT-4 and LLaMA to extract key narrative sentences and assign thematic labels to posts, enabling structured analysis of large corpora of personal narratives[51]. Similarly, hybrid approaches combine topic modeling with LLM interpretation to identify narrative shifts in news discourse[30].

2.3.4 Narrative Aggregation. Taken together, topic modeling and event-based narrative extraction capture complementary aspects of narrative structure. Topic models identify recurring themes across a corpus, while event-based methods reconstruct the temporal progression of actions and actors within texts. Combining these approaches enables the detection of higher-level narrative patterns across many documents, providing a methodological foundation for identifying metanarratives.

While there is no work explicitly identifying metanarratives computationally, there is work where narratives or narrative elements are tracked longitudinally or aggregated in some way. One strand tracks elements as they evolve. For example, de Wildt et al. [19] use probabilistic topic models to trace latent values across large corpora of scientific, policy, and news texts over several decades, showing that values such as sustainability can be detected before they are explicitly named. Rieger et al. [46] combine dynamic topic modeling with temporal and spatial information to model socio-environmental narratives over time. Finally, Reagan et al. [42] analyze large collections of stories and show that the emotional trajectories of narratives tend to follow a limited number of recurring shapes.

Another strand focuses on large scale narrative structure. Norambuena et al. [37] introduce narrative maps, an algorithmic representation that connects events into a directed structure capturing how a story unfolds, providing a route from isolated events to coherent trajectories. Building on event-level representations, Antoniak et al. [6] model narrative paths in birth stories by tracking topic progressions across narrative time and combining them with a verb-based power and agency lexicon, revealing how power relations shift over the course of the story.

Taken together, these studies demonstrate that narrative patterns can be detected by aggregating structured elements such as emotions, events, or actors across large corpora. The resulting patterns typically describe recurring narrative structures within stories rather than overarching narrative trajectories that organize multiple texts. Metanarrative modeling can be an extension of these works, where it aggregates thematic and narrative structures across texts over time in order to identify recurring large-scale patterns.

2.4 Research Gap

Despite increasing interest in computational analyses of fashion discourse, several important gaps remain. Conceptually, narratives within fashion texts have received limited systematic study. While fashion is widely understood as a cultural communication system, it remains unclear how narratives manifest in fashion reviews and how higher-level structures such as metanarratives can be defined within this domain.

Methodologically, metanarratives have not been inferred computationally. The single quantitative study of metanarratives relies on manual coding and collapses each component into its most salient type [28]. Prior work does combine thematic and narrative structure [6], or longitudinal analysis of components [42, 46], which provides a foundation for identifying metanarratives. However, these approaches address separate issues, where a combination of them would be the strongest case for a computationally identifiable metanarrative.

Moreover, none of this work has addressed the fashion domain. Computational analysis on fashion texts usually concerns fashion trend detection or recommender systems [8, 20, 21]. It also remains open how narrative elements in fashion data can be identified and annotated at scale.

Addressing these gaps is necessary to move fashion analytics beyond trend detection and infer cultural meaning computationally by identifying (meta)narratives in fashion discourse.

3 Methodology

3.1 Metanarrative Operationalization

A runway review describes two things: the collection itself, which is a matter of garments, fabrics, and colors, and the show as an event, in which a designer does something with those materials and a critic judges the result. In order to capture metanarratives, both of these need to be mapped and checked for recurring structures, to see if a higher-level representation of these components is possible within the framework of actors, actions, and themes as discussed in the section above. Therefore, two approaches will be combined on the level of an individual review. Topic modeling will be applied to capture *themes* using nouns and adjectives, and event detection will be applied to capture *actions* using verbs. Both strands are developed and investigated at the review level first, and then aggregated to the corpus level in Section 3.6. The designer is taken as the *actor*, since each review is organized around exactly one fashion house whose collections recur across seasons.

3.2 Dataset and Cleaning

The dataset consists of 6,629 runway reviews accompanied by metadata including `year`, `season`, `designer`, `author`, `city`, `date`, `image`, `review`. The reviews span from 1999 to 2014. Note that fashion shows happen half a year before their season, meaning that the corpus covers shows from Spring 2000 to Fall 2014 in fashion seasons.

To ensure a clean dataset, standard text preprocessing was applied prior to analysis. Duplicate, incomplete, and filler entries were removed. Encoding errors were removed using custom-built dictionaries for `designer`, `review`, and `author`. Redundant entries were removed from the `city` and `author` columns, leaving 6,501 reviews with columns `year`, `season`, `designer`, `author`, `city`, `date`, and `review`. The spread of reviews across years is visible in Figure 1.

3.3 Topic Modeling

Topic modeling is applied to detect recurring thematic structures across the corpus over time. It is applied using RollingLDA [44], a sequential extension of Latent Dirichlet Allocation [12], the canonical approach for probabilistic topic modeling. RollingLDA is preferred over a static model due to the assumption that fashion discourse is not constant over the 14-year period of the dataset.

3.3.1 Model and Interpretation. Each review is reduced to a bag of noun and adjective lemmas, from which uninformative and domain-generic terms are removed and the vocabulary is filtered to terms occurring in at least five documents. Against a fashion ontology built from Fashionpedia [27] and manual additions, this confirms the surviving vocabulary is domain-bearing. A RollingLDA with $k = 10$ topics is fitted over one initialization chunk from 1999-2003 with eleven annual update chunks. All ten topics return mean inter-chunk cosine similarity above 0.87, so a single global label per topic is warranted. Topics are interpreted using relevance [52] at $\lambda = 0.4$, which raises the exclusivity of the mean top-15 words from 0.35 to 0.96. Full preprocessing, ontology construction, vocabulary filtering, initialization, fitting specification, and the relevance treatment are discussed in Appendix A.

3.3.2 Topic Validation and Annotation. Construct validity needs to be verified. Furthermore, as this study aims to investigate topics over time, qualitative assignment is preferred over model assignments as semantic validity cannot be derived from a word distribution alone [34]. Therefore, a coding study was conducted using a codebook developed from the topic modeling outputs.

Codebook. The codebook was constructed from four streams of model evidence: the top terms by raw probability and by relevance ($\lambda = 0.4$), both overall and per year, the ten most representative documents according to the model assignment θ, and the chunk-to-chunk stability of each topic. Each topic was given a label together with explicit inclusion and exclusion criteria. The full codebook can be found in Appendix B.

Coding study. Sample size was decided based on a precision floor and the annotator's available time. Inverting the asymptotic variance of Cohen's κ [17], $N_{\text{floor}} = z^2 p_0(1-p_0)/[w^2(1-p_e)^2]$, with $p_e = 0.1017$ from the empirical dominant-topic distribution, $p_0 = 0.61$ [29], $w = 0.10$ and $z = 1.96$, gives $N_{\text{floor}} \approx 113$. A sample size of $N = 200$ was selected, to sit comfortably above this precision floor. A sample of 200 reviews was drawn, stratified by year in proportion to the corpus year distribution. Each review was presented in full, and two human coders independently assigned one topic according to the codebook, a single-label forced-choice task.

An LLM served as a third annotator on the same codebook, both to validate its agreement with the human annotators and to enable efficient annotation of the full dataset, following evidence that LLMs can match trained human annotators on codebook-driven classification [24, 58]. The LLM provenance can be found in Appendix B.1. Agreement is reported as pairwise Cohen's κ since human-human and human-model agreement are different quantities.

Table 1: Pairwise Cohen's κ between annotators and the model's dominant-topic assignment ($N = 200$, ten nominal categories).

Pair	Type	κ
Coder 1 × Coder 2	human–human	0.660
Coder 2 × LLM	human–LLM	0.332
Coder 1 × LLM	human–LLM	0.330

3.3.3 Review Level Analysis. Where the model supplies a distribution over the ten topics, each review is assigned to exactly one topic according to the codebook. This decision was reproducible between human coders. Human-human agreement in the sample study reaches $\kappa = 0.660$, substantial on the benchmark of Landis and Koch [29], whereas both human-LLM pairs fall in the fair range at $\kappa \approx 0.33$ (Table 1). LLM-human agreement falls short of human-human agreement, and this is carried forward as a limitation on the corpus-wide labels. Provenance for these can be found in Appendix B.2. The thematic dimension of metanarrative is measurable on a single review, which is the precondition for aggregating it. Whether these reviews aggregate to corpus-level patterns is taken up in Section 3.6.

3.4 Event Detection

Events are extracted from each review by adapting the event representation of Chambers and Jurafsky [16], which defines the narrative unit of event chains as partially ordered sets of events sharing a common protagonist that is induced from the text, and recovered through coreference resolution. Here, the protagonist is supplied by the metadata (Section 3.2): each review is associated with exactly one designer, and every event in a review belongs to that designer's chain by construction. This was done keeping in mind that to generalize to metanarratives, the designer is the consistent protagonist when aggregating to multiple reviews. What is retained is the event itself, and the sequence in which they appear. The partial-order structure is treated as a hypothesis instead of an assumption here: Section 3.4.2 measures the information carried by event order, which refers to *event sequences* rather than *event chains* to mark the departure from Chambers and Jurafsky [16]. Concretely, dependency parsing via spaCy [26] is used to extract verb-argument triples from each sentence. A triple (subject, verb, object) is saved for each verb that is not an auxiliary attached to another verb and that carries a grammatical subject. The direct object may be empty. Each triple stores the subject, verb, and object lemmas along with the sentence index and a passive voice flag. This yields, for each review, an ordered set of events that may be

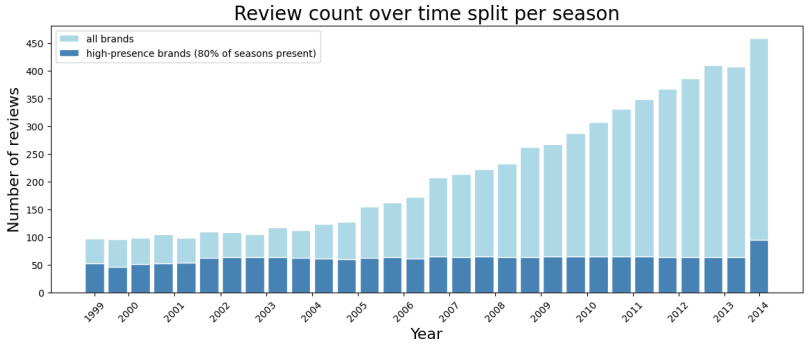


Figure 1: Review count per year, split into seasons (2 per year). The subset of reviews that belong to the designers that are present for more than 80% of the fashion seasons is highlighted separately.

characterized either by their order or by their composition. Both are tested below.

3.4.1 Verb Clustering. In order to make the verb representations less sparse, raw verb lemmas are abstracted into a fixed set of domain event types through k-means clustering of verb embeddings produced by a sentence-transformer model (sentence-transformers/all-mpnet-base-v2) [43, 54]. A distributional representation is preferred to a hand-built lexical resource such as FrameNet [7], as the fashion criticism vocabulary (e.g. "drape", "cinch", "layer", "embroider") is not sufficiently covered by general-purpose inventories. Following diagnostic sweeps over cluster counts from $k = 5$ to $k = 100$, within-cluster coherence remained around the same value for the sweep (mean silhouette $\approx$ 0.03-0.04), providing no quantitative basis for selecting k. The final solution retains $k = 30$ clusters, as a qualitative inspection of the different granularities finds that 30 clusters lead to the most interpretable and distinct results. These clusters are then grouped into two different parent types. One is based on *agency*, which assumes a verb will represent either the stance of the critic or an aspect of the designer or collection. Ambiguous clusters are marked as such. Another is based on a study of book reviews [36], which splits the verbs into four categories: introducing, outlining, highlighting, and evaluating. The full labeling process and the label overview for all three alphabets can be found in Appendix C.

3.4.2 Order. To test whether order within reviews carries information, two quantities were measured: the mutual information between consecutive events, normalized by the entropy of the following event, and each event type's mean relative position. The null shuffles each review's event sequence in place, destroying only ordering, so the observed value is read against what that same review would produce by chance. The decision rests on magnitude rather than the p value, as 119,377 total events can produce significant dependencies that are not informative. To report this, effect size is set as the fraction of successor entropy explained (for the dependency) and Cramér's V against Cohen's 0.10 benchmark (for the positional spread)[18]. To account for order beyond directly consecutive events, the dependency is recomputed at skip-k event gaps to account for event distance.

3.4.3 Composition. To test for composition, each review is reduced to a profile where the proportion of its extracted events falling into each event type. The default for measuring similarity in text-based work is the vector space model, comparing documents by the cosine of the angle between weighted term vectors [48]. Cosine is well motivated in information retrieval, but angle- and Euclidean-based measures applied to closed data are subcompositionally incoherent, meaning that a comparison between two reviews depends on which event types are included [1]. Seeing as the median chain length is 17, a review is not likely to cover every event type, so any comparison between reviews is made over a subset of the full scheme. A similarity measure should therefore give the same verdict about two reviews regardless of which event types are included in the comparison [2, 39].

A second issue is that a similarity value is uninterpretable without a reference distribution. Permutational multivariate analysis of variance supplies this by partitioning variation with respect to an arbitrary distance and generating the null by permutation [4, 5]. A complementary multinomial null is informative because event counts in text are characteristically overdispersed rather than independent [33].

Composition was therefore analyzed in log-ratio geometry (Aitchison distance), and inference was made against explicit permutation and multinomial nulls rather than read off an uncalibrated similarity score.

3.5 Review-Level Analysis

In order to check whether composition carries signal, the review-level profile is checked. For each event type and both of their parent types, the observed variance is compared against a multinomial null in which every review is redrawn from a single corpus-wide distribution at its own chain length. The statistic is the overdispersion index, the ratio of observed to null variance; a value near 1.0 indicates that the observed spread of profiles is simply sampling noise, a value over 1.20 indicates between-review variation beyond what sampling produces. The latter threshold is calibrated so that a true SD of 5 percentage points is cleared on the dominant event type on all three alphabets.

3.6 Corpus-Level Analysis

The preceding sections establish what themes and actions can be measured on a single review: one thematic label, and one event-type profile whose dispersion exceeds what sampling alone would produce. In order to look for metanarratives, the combination of *themes* (topics), *actions* (events) and *actors* (the designer) will be considered across the span of years of the corpus.

In order to analyze how metanarratives evolve over the corpus, the paper by Kaplan et al. [28] looked at the individual components of metanarratives as well as the combination of them across documents in their corpus. The question is whether reviews sharing a designer or a period of time resemble one another more than reviews drawn at random. Because the topics and events per review yield a different type of object (a single nominal topic label and a compositional vector for actions) they are aggregated separately before being considered together.

3.6.1 Themes. To model thematic prevalence, the annotation of the 6,501 reviews in the corpus was used.

For every topic, a logistic regression is fitted in which the outcome is an indicator for whether a review was assigned to topic k, and the predictor of interest is the publication year, with a fixed effect for season (Fall/Spring). Standard errors were clustered by designer. All ten topics were tested simultaneously, and p-values were adjusted for the false discovery rate using the Benjamini-Hochberg procedure, and significance is assessed at $q < 0.05$.

Two further tests introduce the designer: each topic's mean relative position within a designer's show-ordered trajectory is summarized by Cramér's V against Cohen's 0.10 benchmark, and the dependence of the topic label on the designer is tested against a topic-label permutation null, again with Cramér's V as the effect size net of the association any partition yields by chance.

3.6.2 Actions. The event-type profile of a single review is a composition: a vector of proportions carrying only relative information. Comparisons between such profiles therefore require the log-ratio geometry motivated in Section 3.4.3. Counts are centred-log-ratio (CLR) transformed,

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right], \qquad g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \quad (1)$$

so that ordinary Euclidean distance in the transformed space equals Aitchison distance.

Group differences are assessed by PERMANOVA [4]. The statistic is $\eta^2 = \mathrm{SS}_{\mathrm{between}}/\mathrm{SS}_{\mathrm{total}}$, the share of compositional variance attributable to the grouping, and the reference distribution is obtained by permuting group labels while preserving group sizes.

Raw η^2 is not interpretable on its own, since any partition of n observations into g groups explains approximately $(g-1)/(n-1)$ of the variance by construction a baseline ranging from 0.002 to 0.494 across the groupings in Table 2. All results are therefore reported as the excess over the permutation null,

$$\eta^2_{\mathrm{excess}} = \eta^2_{\mathrm{obs}} - \bar{\eta}^2_{\mathrm{null}}. \quad (2)$$

.

Tier 1: single groupings. Reviews are grouped by year, by author, and by designer in turn. These groupings separate three sources of structure that could organize the corpus: a period effect, in which the way collections are written about changes over time; an author effect, and a designer effect.

Table 2: Experimental design for the composition analysis. Tiers 1 and 3 share one fixed panel of the 65 designers with at least 24 reviews each (1,883 reviews); Tier 2 minimum cell sizes were fixed in advance from the distribution of cell sizes.

Tier	Reviews grouped by	Min.	Groups	Reviews
1	year	24	16	1,883
1	author	24	36	1,883
1	designer	24	65	1,883
2	designer × 3 years	6	518	3,128
2	designer × year	2	2,649	5,319
3	within designer	24	65	1,883

Tier 2: designer over time. Grouping by designer × year, and by designer × three-year window, tests whether a house's compositional profile is specific to a moment in its trajectory rather than fixed across its coverage. The three-year window is the coarser of the two and trades resolution for larger cells; the annual grouping admits cells of two reviews and is correspondingly noisier.

Tier 3: within designers. To test whether a designer's reviews are internally coherent, the mean pairwise Aitchison distance among its own reviews is compared against random review sets of the same size, giving a measure of internal coherence that does not depend on how that designer compares to any other. Temporal drift within a designer is then measured as the Spearman correlation between the temporal separation of two reviews and their compositional distance, with a null obtained by permuting years within the designer, so that the designer's own dispersion is held fixed and only the temporal ordering is destroyed. A designer may thus be coherent and static, coherent and drifting, or neither, and these are distinguished.

3.6.3 Synthesis. With the designer as the common actor, the roles themes and actions are considered across the corpus. They are however not combined into a single measure, as a joint construct would require both dimensions to be established on the same unit, and the compositional profile of events does not support this. Themes and actions are therefore reported alongside one another rather than tested as one quantity.

4 Results

Following the metanarrative definition in Section 3, the results are reported for topics and events, representing themes and actions, and for their interactions with the designer, representing the actor. Reporting the results along these dimensions addresses sub-question 1, by establishing how metanarrative components can be operationalized in the fashion domain. For the analyses that include the designer, a minimum review threshold is imposed so that a designer is observed often enough to estimate a stable profile and to account for one-time reviews acting as noise, ensuring a comparable dataset. Requiring at least 24 reviews, one per season over

80% of the corpus, leaves 65 of the 709 designers (9.2% of designers, 29.2% of all reviews). When considering the author, empty-author reviews were not included in the analysis, dropping 12 reviews.

4.1 Topics

This section reports the thematic dimension of the operationalization, addressing the second sub-question, from the side of *themes*: topics and their occurrence in reviews as building blocks of a meta-narrative. The RollingLDA model yields ten topics: each review was assigned a single dominant topic. Based on the outputs of the RollingLDA, the ten topics are classified as follows:

(0) *masculine-inflected tailoring and leather*
(1) *glamour, eveningwear and occasion dressing*
(2) *graphic print and nightlife dressing*
(3) *luxury fabrics, furs and outerwear*
(4) *designer vision, technical mastery and intellectual authority*
(5) *conceptual staging and theatrical collection*
(6) *lace, lingerie and sensuality*
(7) *sportswear, denim and youth casualwear*
(8) *feminine spring lightness and romantic delicacy*
(9) *utilitarian separates and wearability*

Which topics are most prevalent? Using the annotated labels, thematic prevalence is uneven, ranging from 4.0% to 14.5% of reviews. The most common themes are graphic print and nightlife dressing (topic 2, 14.5%), lace and lingerie (topic 6, 13.7%) and utilitarian separates (topic 9, 12.4%); the least common are designer vision (topic 4, 4.0%) and feminine spring dressing (topic 8, 5.1%). Topic prevalence is shown per topic, per year in Figure 2 (see Appendix D Figure 7 for a combined overview). Topic prevalence differs per season, which is shown in Figure 8 in Appendix D.

4.1.1 Topics Structure. Introducing the *actor* into the thematic dimension, this analysis addresses the third sub-question, by asking whether themes are organized by the designer as a protagonist and whether this shifts over time. For all analyses, the subset of 65 designers with at least 24 reviews was used, as these have 80% coverage over time across the corpus.

Does prevalence of topics change over time? For each topic a logistic regression of the label indicator on calendar year (with a season fixed effect and designer-clustered standard errors) was fitted on the subset of 65 designers, and p-values were adjusted across the ten topics with Benjamini-Hochberg. No topic changes significantly at $q < 0.05$: the strongest movement is a decline in utilitarian separates (topic 9, $OR = 0.962$ per year, $q = 0.056$) which falls just short of the threshold.

Do topics have a fixed place in a designer's trajectory over the corpus? Taking each designer's reviews in show-date order, each topic's mean relative position within the trajectory was measured, mirroring the positional test of the event analysis (Section 3.4.2) on an inter-review level this time. A faint arc surfaces only among the 65 most heavily covered designers (≥ 24 reviews; range 0.172, Cramér's $V = 0.095$) and sits below Cohen's small effect threshold of 0.10.

Is there a characteristic topic composition per designer? Grouping the corpus by designer, the dependence of the topic label on the designer was tested against a whole corpus baseline, with Cramér's V as the effect size and a topic-label permutation null subtracting the association any partition yields by chance. Across the 65 designers with at least twenty-four reviews (1,897 reviews), observed $V = 0.340$ far exceeds the permutation-null mean of 0.184, an excess of +0.156 ($z = 27$, permutation $p \leq 0.005$).

4.2 Events

This section reports the event dimension, addressing the second sub-question from the side of *actions*: the recurring event structures that reviews draw on, tested for whether their order or composition generalizes across the corpus. Events consist of verbs, with each verb assigned to one of the 30 event types by a transformer model as discussed in Appendix C. These contain 2 variations of parent types: rhetorical moves, which outline the section of a review [36], and agency, which assigns the verbs to either the critic, the designer or collection, or leaves them as ambiguous. Event mass is highly concentrated in all three alphabets. For the 30 clusters, the single largest type (*being & presentation*) accounts for 30.4% of all events, covers roughly 46% of the corpus with *generic action* (15.7%); the tail runs down to *refinement* at 0.3%. Figure 4 shows chain length distribution for individual reviews, and Figure 3 shows the composition of individual reviews.

Does order matter within a single review? Testing event order against a within-review shuffle null, both quantities fall short of their effect-size thresholds on the 30-type alphabet. The dependency between consecutive events exceeds chance by only 0.1% of the successor entropy, and this decays to near zero by a gap of five events. Mean relative position is also flat (positional spread = 0.105). Cramér's V equals 0.03, below Cohen's 0.10 benchmark. The coarser alphabets are even weaker. Knowing the current event says almost nothing about the next, ordering was not pursued and composition was carried forward.

Does composition matter within a single review? Reducing each review to a profile of event-type proportions and comparing its spread against a multinomial null (each review drawn from one corpus-wide distribution at its own length), the overdispersion index sits at ≈ 1.0 at every granularity (1.04 for the 30 types, 1.01 for the rhetorical parents, 1.05 for the agency parents), all below the 1.20 floor. At the level of a single review, event composition is indistinguishable from sampling noise.

4.2.1 Events × Designer. Because review-level profiles carry no signal, the question is whether aggregating reviews does, and if so whether this holds up over time or per designer, or both. Composition was analyzed in log-ratio (Aitchison) geometry via PERMANOVA, with results reported as the excess of observed η^2 over the permutation-null baseline (η^2_{excess}); Table 3 gives the fine (30-type) alphabet. To decide on significance, η^2_{excess} was set as a threshold rather than a p value, as the size of the corpus at 6501 reviews is large enough that even a small deviation from the mean would result in a significant test. Therefore, the design tests against an effect size floor of 0.02, as this is a common threshold for a weak effect size. Author is included as a control. Assumptions are checked [1]. This

[1] PERMANOVA's three assumptions are checked in notebook 3.4 (see Github link); none holds strictly, but the design is robust to each.

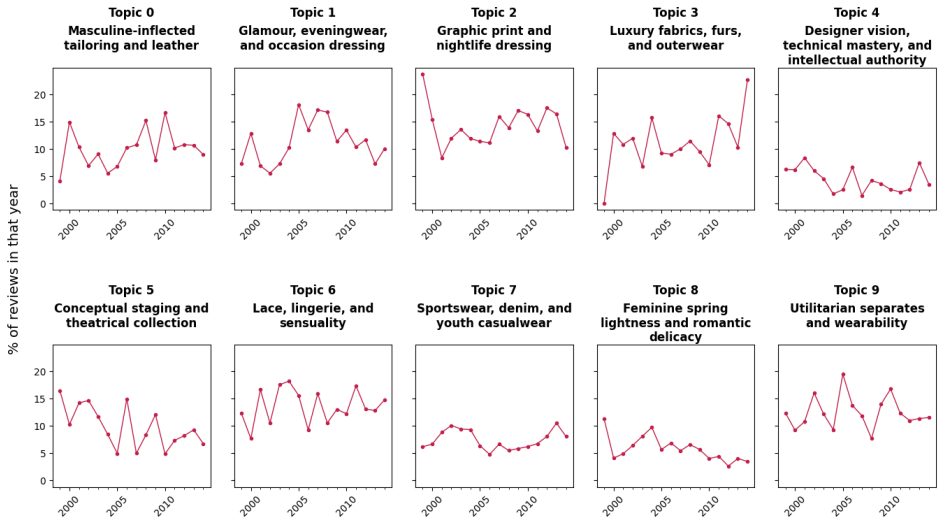


Figure 2: Topic share per year, shown for all 10 topics individually.

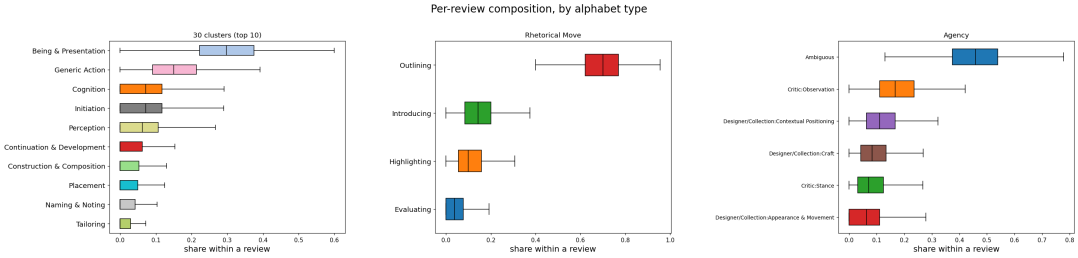


Figure 3: Event composition per alphabet over time. The 30 cluster mapping shows the top 10 most common events within reviews.

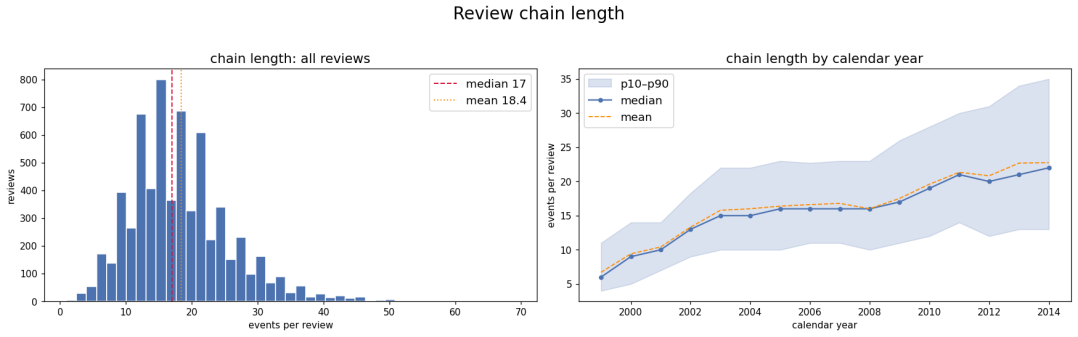


Figure 4: Chain length for the whole corpus and per calendar year.

analysis addresses the third sub-question by testing whether event composition is organized by designer and evolves across years.

Tier 1: single factors. The designer grouping on the 65 most common designers fails the floor at every alphabet (+0.012 on the 30-cluster), while the author clears it at the 30-cluster as well the rhetorical resolutions +0.034, +0.022). On the 30-cluster alphabet, the author's excess nearly triples the designer's. The year clears the floor on the 30-cluster alphabet (+0.030), meaning that the composition shifts across calendar time, but this decays on the coarser alphabets.

Tier 2: designer × time. A designer signal surfaces only once a designer is pinned to a moment in time (designer × 3yr, +0.026). The designer × year cell is structurally uninterpretable and reported for completeness, with only two reviews per designer-year, each

Table 3: Event composition by grouping (30-type alphabet). η^2_{excess} is observed minus null-baseline variance explained; bold marks groupings that clear the 0.02 effect-size floor.

Tier	Grouped by	Groups	Reviews	η^2_{excess} 30-type	MR (4)	Agency (6)
1	year	16	1,883	**+0.030**	+0.010	+0.014
1	author	36	1,883	**+0.034**	**+0.022**	+0.017
1	designer	65	1,883	+0.012	+0.015	+0.019
2	designer × 3yr	518	3,128	**+0.026**	+0.014	**+0.023**
2	designer × year	2,649	5,319	+0.036	+0.032	+0.025
3	within-designer (consist.)	65	1,883	−0.017	+0.001	+0.020
3	within-designer (drift)	65	1,883	**+0.064**	−0.005	−0.001

group is so small that a purely random partition already explains half the variance (its null baseline is 0.498 by construction).

Tier 3: within a designer. Composition is no more consistent than random review sets, but does drift slowly over its own history on the fine alphabet. This effect is absent on both of the coarser alphabets.

5 Discussion

5.1 Detecting metanarratives computationally in fashion reviews (SQ1)

The operationalization of Section 3 maps three distinct components of a metanarrative (actors, actions, themes). Substituting a descriptive theme for Kaplan et al.'s *value component* [28] follows directly from the argument in Section 2. The contribution is that these components are inferred computationally at corpus scale rather than hand-coded and collapsed to a single most salient type, which is the methodological gap defined in Section 2.

The two components differ as objects, and this difference carries through to the results. The thematic component is separable per review: RollingLDA returns ten topics, where each review is assigned a single label, and every topic is populated after annotation. The extracted events collapse into a heavily concentrated alphabet, where *being & presentation* and *generic action* cover close to half the corpus, with a long thin tail. A theme is a discrete choice whereas an action is a proportion.

5.1.1 *Recurring structures (SQ2).*

Themes. Prevalence is uneven but never absent. Figure 2 in Section 4 shows that most topics fluctuate from year to year without a directional trend. The seasonal split shows that furs and outerwear (Topic 3) concentrate in Fall, while feminine lightness, lace, and graphic nightlife dressing (Topics 6, 8, and 2) concentrate in Spring. This matches the seasons content wise, as lighter fabrics and nightlife are likely to be included in the Spring/Summer season, and heavier fabrics more so in Fall/Winter seasons.

Actions. Event mass is concentrated rather than distributed (Figure 4 in Section 4).

Order. No recurring sequential structure survives. While the full approach for event chains with coreference resolution [16] was not adopted, events were clusters were assigned a parent label based on agency, the designer was used as a protagonist for topic models. Within a review, one event does not depend on the events that came before it, and topics have no characteristic positions across the reviews of a specific designer. The assumption that events fall into a recoverable order around a protagonist was not met for topic models across designers. On the review level, the lack of order of events is not organized around a protagonist due to the length, but no order can be found regardless.

5.2 Evolution across designers and years (SQ3)

Aggregating supplies the signal that single reviews lack, and it splits cleanly by component. On themes, the designer is decisive: topic composition depends on the house far beyond chance ($V = 0.340$ against a permutation-null mean of 0.184). This is the strongest evidence for a metanarrative, since the designer is the one protagonist across seasons. However, this thematic identity is static, as no topic trends significantly over calendar time.

On actions the actor changes. Under PERMANOVA against the 0.02 excess-variance floor, the designer grouping fails, while the author clears it. The year clears it on the fine alphabet, meaning composition shifts measurably across calendar years. This means that event composition is a fingerprint of the critic and the period, not of the designer. A designer-specific signal appears only once a designer is pinned to a three-year window.

5.3 Limitations

There are several limitations to these findings. First, the effect sizes reported here are consistently small, meaning that the measured effects should not be over-interpreted as strong predictions.. On reliability and validity, corpus-wide topic labels rest on an LLM annotator whose agreement with humans is only fair ($\kappa \approx 0.33$ against human-human 0.660), so the thematic results carry that uncertainty. Regarding events, no coreference resolution takes place, limiting the claim on the lack of event order for individual reviews. On construct validity, the dominance of *being & presentation* reflects the copula absorbing too much of the event mass, so the validity of *action* is thinner than that of *theme*. On generalizability, the method can be adapted to different corpora discussing fashion, such as social media, but the results are likely to differ strongly, as the outcomes are quite specific to fashion reviews in the format of the dataset.

6 Conclusion

This study set out to identify metanarratives in fashion runway reviews computationally. This was organized around three sub-questions: how metanarratives can be detected, what structures emerge, and how they evolve across designers and years.

SQ1: detecting components. A metanarrative was operationalized as three components inferred at corpus scale: the designer as *actor*, RollingLDA topics as *themes*, and events as *actions*. Themes are recoverable as discrete per-review labels, actions are recoverable only as a compositional profile explained by the author of the review, replacing the designer as actor. The components are therefore reported side by side rather than fused.

SQ2: recurring structures. The recurring structure is compositional, not sequential. Order carries no signal, within a review or across a designer's reviews, and a single review's event composition is indistinguishable from sampling noise. Reviews describe far more than they evaluate, and lengthen over the corpus while their composition remains flat. Structure only appears when reviews are aggregated.

SQ3: evolution across designers and years. Aggregation reveals that the designer organizes the thematic composition, while the critic organizes the event composition. What a collection is written about belongs to the fashion house, and how it is told belongs to the reviewer. Only the thematic component meets the designer-as-protagonist criterion of a metanarrative.

Contribution and Future Work. The contribution of this research is to move fashion analysis beyond trend and sentiment detection and to show that a metanarrative can be identified computationally. while separating components on several levels that manual coding would merge, thereby exposing that different components are governed by different actors in the fashion domain. This answer is qualified by its limitations as mentioned in Section 5. Therefore, suggestions for future work include representing actions in a richer manner, taking subjects and verb complements such as predicative adjectives into account. This way, a review can be properly separated between evaluative and descriptive statements, which lead to a more accurate representation of the narrative component of *action* as well as *actor*. Furthermore, annotation on topics over time should be either done manually, or including an LLM that is better suited for the fashion domain to ensure topics are labeled accurately. With the improved event and topic representations, the study could extend to corpora beyond fashion reviews, and could identify metanarratives across textual fashion data from other sources such as social media and designer's descriptions of themselves on their own channels. It could also include different modes, such as images, to track descriptive themes. Lastly, the fashion ontology and the verb mappings are reusable resources: future work on fashion text can extend them into richer representations of the domain's specialized language.

References

[1] J. Aitchison. 1982. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* 44, 2 (1982), 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x

[2] J. Aitchison. 1992. On Criteria for Measures of Compositional Difference. *Mathematical Geology* 24, 4 (1992), 365–379. doi:10.1007/BF00891269

[3] Jorge Ale Chilet, Cuicui Chen, and Yusan Lin. 2017. Runway Reviews and Instagram Activities of High-End Fashion Brands. doi:10.7910/DVN/BCOSKY

[4] Marti J. Anderson. 2001. A New Method for Non-Parametric Multivariate Analysis of Variance. *Austral Ecology* 26, 1 (2001), 32–46. doi:10.1111/j.1442-9993.2001.01070.x

[5] Marti J. Anderson and Daniel C. I. Walsh. 2013. PERMANOVA, ANOSIM, and the Mantel Test in the Face of Heterogeneous Dispersions: What Null Hypothesis Are You Testing? *Ecological Monographs* 83, 4 (2013), 557–574. doi:10.1890/12-2010.1

[6] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 88:1–88:27. doi:10.1145/3359190

[7] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://aclanthology.org/C98-1013/

[8] Aayam Bansal and Agneya Tharun. 2025. *Emotional Analysis of Fashion Trends Using Social Media and AI: Sentiment Analysis on Twitter for Fashion Trend Forecasting*. doi:10.48550/arXiv.2505.00050

[9] Malcolm Barnard. 1996. *Fashion as Communication*. Routledge, London. doi:10.4324/9781315013084

[10] Barthes. 1990. *The Fashion System*. University of California Press, Berkeley.

[11] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 113–120. doi:10.1145/1143844.1143859

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[13] Federica Bologna, Rosamond Thalken, Kristen Pepin, and Matthew Wilkens. 2025. Endometriosis Communities on Reddit: Quantitative Analysis. *Journal of Medical Internet Research* 27, 1 (March 2025), e57987. doi:10.2196/57987

[14] Jerome Bruner. 1991. The Narrative Construction of Reality. *Critical Inquiry* 18, 1 (Oct. 1991), 1–21. doi:10.1086/448619

[15] Elizabeth Castaldo Lundén. 2020. Exploring Fashion as Communication: The Search for a new fashion history against the grain. *Popular Communication* 18, 4 (Oct. 2020), 249–258. doi:10.1080/15405702.2020.1854952

[16] Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui (Eds.). Association for Computational Linguistics, Columbus, Ohio, 789–797. https://aclanthology.org/P08-1090/

[17] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[18] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge, New York. doi:10.4324/9780203771587

[19] T. E. de Wildt, I. R. van de Poel, and E. J. L. Chappin. 2022. Tracing Long-term Value Change in (Energy) Technologies: Opportunities of Probabilistic Topic Models Using Large Data Sets. *Science, Technology, Human Values* 47, 3 (May 2022), 429–458. doi:10.1177/01622439211054439

[20] Yujuan Ding, Zhihui Lai, P.Y. Mok, and Tat-Seng Chua. 2023. Computational Technologies for Fashion Recommendation: A Survey. *ACM Comput. Surv.* 56, 5 (Nov. 2023), 121:1–121:45. doi:10.1145/3627100

[21] Yujuan Ding, Yunshan Ma, Lizi Liao, Wai Keung Wong, and Tat-Seng Chua. 2022. Leveraging Multiple Relations for Fashion Trend Forecasting Based on Social Media. *IEEE Transactions on Multimedia* 24 (2022), 2287–2299. doi:10.1109/TMM.2021.3078907

[22] Mai Elshehaly, Mirela Reljan-Delany, Jason Dykes, Aidan Slingsby, Jo Wood, and Sam Spiegel. 2024. Visualising Lived Experience: Learning from a Master Narrative Framing. In *2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 5–14. doi:10.1109/BELIV64461.2024.00006

[23] Équipe de recherche Fabula. 2026. Cultural Narratologies: Form and Context in Contemporary Narrative Theory (Between, vol. XVI.32, Nov. 2026). https://www.fabula.org/actualites/132255/cfp-between-xvi-32-november-2026-cultural-narratologies-form-and-context-in-contemporary-narrative-theory.html

[24] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[25] Gozde Goncu Berk. 2023. Chapter 6 - Trend Lifecycles and Adopter Categories. In *Communicating Fashion: Trend Research and Forecasting*. University of Minnesota Libraries. https://open.lib.umn.edu/communicatingfashion/chapter/chapter-6-trend-lifecycles-and-adopter-categories/

[26] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. doi:10.5281/zenodo.1212303

[27] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge J. Belongie. 2020. Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset. *CoRR* abs/2004.12276 (2020). arXiv:2004.12276 https://arxiv.org/abs/2004.12276

[28] Yael R. Kaplan, Tamir Sheafer, and Shaul R. Shenhav. 2022. Do we have something in common? Understanding national identities through a metanarrative analysis.

Nations and Nationalism 28, 4 (2022), 1152–1172. doi:10.1111/nana.12872

[29] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[30] Kai-Robin Lange, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch. 2025. Narrative Shift Detection: A Hybrid Approach of Dynamic Topic Models and Large Language Models. arXiv:2506.20269 (2025). doi:10.48550/arXiv.2506.20269 arXiv:2506.20269 [cs].

[31] Jian Ming Luo, Huy Quan Vu, Gang Li, and Rob Law. 2020. Topic modelling for theme park online reviews: analysis of Disneyland. *Journal of Travel Tourism Marketing* 37, 2 (Feb. 2020), 272–285. doi:10.1080/10548408.2020.1740138

[32] Natalya Lusty. 2021. Fashion futures and critical fashion studies. *Continuum* 35, 6 (Nov. 2021), 813–823. doi:10.1080/10304312.2021.1993568

[33] Rasmus E. Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*. ACM Press, Bonn, Germany, 545–552. doi:10.1145/1102351.1102420

[34] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures* 12, 2–3 (April 2018), 93–118. doi:10.1080/19312458.2018.1430754

[35] Maryam Molazadeh. 2023. Fashion Semiology based on Roland Barthes' Theory of Layered Semiotics. 2, 2 (Feb. 2023), 271–298. doi:10.22077/islah.2023.5829.1180

[36] Désirée Motta-Roth. 1996. Same Genre, Different Discipline: a Genre-Based Study of Book Reviews in Academe. *The ESPecialist* 17, 2 (1996). https://revistas.pucsp.br/index.php/esp/article/view/9684

[37] Brian Keith Norambuena and Tanushree Mitra. 2021. Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 228:1–228:33. doi:10.1145/3432927

[38] Devi Sari Panggabean and Rahmadsyah Rangkuti. 2020. The Basic Concept of Narratology and Narrative. *Language Circle: Journal of Language and Literature* 14, 2 (2020), 193–202.

[39] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. 2015. *Modeling and Analysis of Compositional Data.* John Wiley & Sons, Chichester, UK. doi:10.1002/9781119003144

[40] Anne Peirson-Smith and Ben Peirson-Smith. 2020. Fashion archive fervour: the critical role of fashion archives in preserving, curating, and narrating fashion. *Archives and Records* 41, 3 (Sept. 2020), 274–298. doi:10.1080/23257962.2020.1813556 _eprint: https://doi.org/10.1080/23257962.2020.1813556.

[41] Andrew Piper. 2023. Computational Narrative Understanding: A Big Picture Analysis. In *Proceedings of the Big Picture Workshop*, Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, and Noah A. Smith (Eds.). Association for Computational Linguistics, Singapore. doi:10.18653/v1/2023.bigpicture-1.3

[42] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (Dec. 2016), 31. doi:10.1140/epjds/s13688-016-0093-1 arXiv:1606.07772 [cs].

[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 (Aug. 2019). doi:10.48550/arXiv.1908.10084 arXiv:1908.10084 [cs.CL].

[44] Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2021. RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2337–2347. doi:10.18653/v1/2021.findings-emnlp.201

[45] Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2024. LDAPrototype: a model selection algorithm to improve reliability of latent Dirichlet allocation. *PeerJ Computer Science* (2024). doi:10.7717/peerj-cs.2279

[46] Jonas Rieger, Felipe Muñoz, Lars Grönberg, Kai-Robin Lange, Iván Ojeda-Pereira, Darío Briceño, Christian Nass, Carsten Stahl, José Cassola, Carolina Rojas-Córdoba, Brian Keith-Norambuena, Marcelo Lufin, Fernando Campos-Medina, and Sebastián Herrera-León. 2026. Following Socio-Environmental Conflict Narratives About Energy Transition in Chile: A Spatio-Temporal Analysis Using Dynamic Topic Modeling. In *Proceedings of the Text2Story Workshop (Text2Story'26)*. https://www.di.ubi.pt/~jpaulo/Text2Story2026/paper5.pdf

[47] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*. 399–408. doi:10.1145/2684822.2685324

[48] G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 11 (1975), 613–620. doi:10.1145/361219.361220

[49] Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review* 56, 8 (Aug. 2023), 8393–8435. doi:10.1007/s10462-022-10338-7

[50] Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, Valencia, Spain, 432–436. https://aclanthology.org/E17-2069/

[51] Mohammad Shokri, Emily Klapper, Jason Shan, and Sarah Ita Levitan. 2025. Finding Common Patterns in Domestic Violence Stories Posted on Reddit. In *Proceedings of the The 7th Workshop on Narrative Understanding*, Elizabeth Clark, Yash Kumar Lal, Snigdha Chaturvedi, Mohit Iyyer, Anneliese Brei, Ashutosh Modi, and Khyathi Raghavi Chandu (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 15–25. doi:10.18653/v1/2025.wnu-1.6

[52] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Jason Chuang, Spence Green, Marti Hearst, Jeffrey Heer, and Philipp Koehn (Eds.). Association for Computational Linguistics, Baltimore, Maryland, USA, 63–70. doi:10.3115/v1/W14-3110

[53] Frances L. M. Smith and Debbie S. Dougherty. 2012. Revealing a Master Narrative: Discourses of Retirement Throughout the Working Life Cycle. *Management Communication Quarterly* 26, 3 (Aug. 2012), 453–478. doi:10.1177/0893318912438687

[54] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv:2004.09297 (Nov. 2020). doi:10.48550/arXiv.2004.09297 arXiv:2004.09297 [cs.CL].

[55] Wikipedia. 2025. https://en.wikipedia.org/w/index.php?title=Lists_of_colors&oldid=1320364460 Page Version ID: 1320364460.

[56] Wikipedia. 2025. https://en.wikipedia.org/w/index.php?title=List_of_textile_fibres&oldid=1277464931 Page Version ID: 1277464931.

[57] Wikipedia. 2026. https://en.wikipedia.org/w/index.php?title=List_of_fabrics&oldid=1361258947 Page Version ID: 1361258947.

[58] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291.

Appendix A Appendix A: Topic Modeling Methodology

A.0.1 Preprocessing for Topic Modeling. Continuing from the cleaning stage, contractions were expanded and all punctuation was replaced with whitespace. The normalized text was then processed with spaCy's (v3.8.14) `en_core_web_lg`, the most accurate of spaCy's CNN-based English pipelines. This was chosen over the transformer pipeline because its marginal tagging advantage does not survive aggregation into a BoW representation, where POS (part-of-speech) tagging functions only as a coarse noun/adjective filter. It was used with named-entity recognition disabled, assigning each token a POS tag and a lemma [26]. From the tagged output, only alphabetic tokens tagged as nouns or adjectives were kept, excluding designer words, spaCy stopwords, words of less than 3 characters. Based on inspection of the dataset, a manual list of uninformative domain-specific words was excluded as well, including synonyms for "designer", uninformative runway-related words, and generic size, time, person, and object indicators (Section A.1). The result is a per-document list of noun and adjective lemmas, resulting in **18,511** unique terms.

A.0.2 Fashion Ontology Building. To track which of these terms constitute fashion vocabulary, a fashion ontology was built using the following three sources: the Fashionpedia ontology [27], the designer names in the `designer` column, and manual additions. The manual additions are in categories considered relevant but not covered by the Fashionpedia dataset: colors, textile fibers, fabric types, gathered from Wikipedia overview lists [55–57], and occasions, style descriptors, and leftover garment and accessory terms identified through inspecting reviews qualitatively. Terms are mapped to categories and subcategories based on Fashionpedia's data in Table 4, which were expanded with the manual additions and the list of designers. All non-designer terms in the ontology were lemmatized using spaCy, deduplicated, and indexed into three lookup sets of unigrams, bigrams, and trigrams before saving. Designer names were left verbatim. Even though designer names are not included in the topic modeling analysis, they were still included in the ontology as this may be useful in other contexts.

Category	Subcategories	Terms
Garment	garment type, garment part, upper-body, lower-body, whole-body, neckline type, waistline, opening type, closure	192
Material	fabric type, textile fibre, non-textile material type	191
Style	style descriptor, silhouette, textile pattern, length, textile technique, embellishment	119
Accessory	head, legs and feet, arms and hands	18
Color	—	22
Occasion	—	5
Brand	—	708
Total		1,255

Table 4: Categories and subcategories of the fashion ontology.

A.0.3 Vocabulary Filtering. In order to filter the vocabulary, the corpus frequency, PPM, TF-IDF, and document frequency proportion are computed for each term. The RollingLDA paper uses a minimum threshold by including words that satisfy a document frequency of 5 as well as a minimum PPM of 10 per chunk, or words that occur more than 100 times in the whole corpus [44]. As this is a much smaller corpus, these metrics are expected to be highly correlated, which is why only the minimum document frequency of 5 across the whole corpus was used as a filter as an initial step. A small sweep was run to see if lowering the threshold would yield a higher retention of fashion terms, but it would at best bring 14 fashion terms along with over 2000 noise words which was seen as an unnecessary tradeoff. No upper bound document-frequency cutoff was applied, as the most frequent surviving term (*jacket*) occurs in only 45.8% of documents, so no term behaves as a corpus-wide stopword. This step reduced the vocabulary to **6,249** terms in total. Measured against the fashion ontology, 366 out of 419 fashion unigrams present in the corpus survive this step, a domain-term retention rate of 87.4%.

A.0.4 Initialization. Before fitting the full model, an initial LDA model is fitted in order to decide on the initialization cutoff and the number of topics K. First, the initialization period was decided based on two requirements: the initial period must be long enough to estimate stable topics and cover enough of the vocabulary, but short enough to leave a substantial rolling period over which change can be observed [44]. The period from 1999 to 2003 containing 936 documents was selected as it covers 75.2% of the vocabulary, which is a majority solid enough to anchor the topics while still reserving the bulk of the corpus (2004-2014) as the rolling period over which new vocabulary can enter. In order to decide the number of topics K, a sweep was conducted over the initial part of the corpus, over a range 5 to 25 topics, with five LDAPrototype models from different random seeds per K. For each K, mean C_v coherence [47] and mean cross-seed cosine similarity were measured, to check topic quality and reproducibility. Cross-seed similarity declined monotonically with K while coherence remained essentially flat, so neither criterion isolated a clear optimum, as seen in Figure 5. $K = 10$ was chosen above the stability optimal $K = 5$ to leave room for topics emerging over the rolling period. The full sweep setup can be found in Table ??.

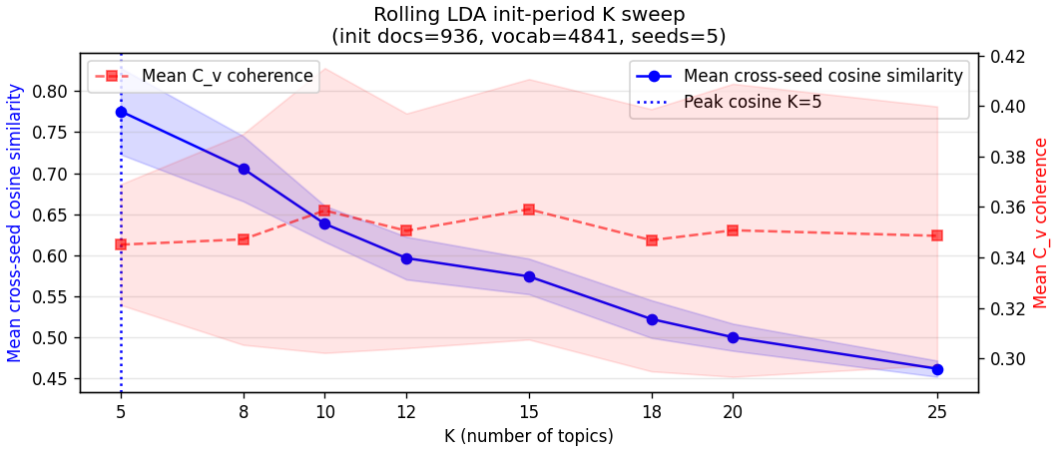


Figure 5: Topic-number sweep on the initialization period (1999-2003); 936 documents with five random seeds per K. Mean cross-seed cosine similarity (blue, left axis) measures how reproducibly topics recur across seeds; mean C_v coherence (red, right axis) measures the semantic coherence of each topic's top words. Shaded bands show ± 1 standard deviation across seeds.

Setting	Decision
Initialization period	1999-09-12 – 2003-12-31 (936 documents)
Vocabulary coverage at cutoff	75.2%
Initial-model estimator	LDAPrototype
Topic counts swept	5, 8, 10, 12, 15, 18, 20, 25
Seeds per topic count	5
Selection criteria	cross-seed cosine similarity; C_v coherence
Chosen number of topics K	10

Table 5: The settings for the sweep on the initialization period.

A.0.5 RollingLDA. Using the set hyperparameters, the final model was fitted with the `rollinglda` package (v0.1.4), the reference implementation of [44], driven from Python through `rpy2`. The initialization period from 1999-2003 was re-fitted using LDAPrototype, this time with 100 candidate runs, which is the default value, and a final prototype selected by pairwise topic similarity and recommended by the RollingLDA paper [44, 45]. The full model was fitted with $K = 10$ topics, symmetric priors $\alpha = \eta = 1/K$, and 200 Gibbs sampling iterations, applied to 6,501 documents. Each chunk contains a year, which contains both the spring/summer and fall/winter runway seasons. A two-year memory window was chosen to reflect the multi-season arc over which fashion trends typically emerge and consolidate [25]. This preserves topic continuity while allowing for the vocabulary to grow. The full corpus is thus modeled as one initialization chunk followed by 11 annual update chunks, 12 in total, over which the vocabulary grows from 1,671 terms to 2,971 terms. The model hyperparameters are given in Table **??** in Appendix A.

Topic-word distributions were inspected under both raw ϕ-ranking ($\lambda = 1.0$) and relevance ranking at $\lambda = 0.4$ [52], the lower value consistent with the initialization findings. All ten topics returned mean inter-chunk cosine similarities above 0.87, with broadly even values across the observation period as seen in Figure 6.

A.0.6 Topic Interpretation. On a domain-homogeneous corpus, ranking a topic's terms by raw probability is uninformative, as a small set of ubiquitous words dominates the top of almost every topic. Topics were therefore interpreted based on relevance [52], which down-weights corpus-frequent terms in favor of terms distinctive to a topic. Following inspection of the topics, λ was set to 0.4, below the $\lambda = 0.6$ rule of thumb, as the lower value produced more interpretable rankings. This is supported qualitatively: the mean exclusivity of the top fifteen terms per topic rises from 0.35 under raw probability to 0.96 under *relevance*

$$r(w \mid k) = \lambda \log p(w \mid k) + (1 - \lambda) \log \frac{p(w \mid k)}{p(w)},$$

at $\lambda = 0.4$. This is also consistent with literature on stopword removal, as it is a common problem that completely removing words that dominate topics results in the next layer of common words dominating the topics [50]. Schofield et al. argue that handling corpus specific

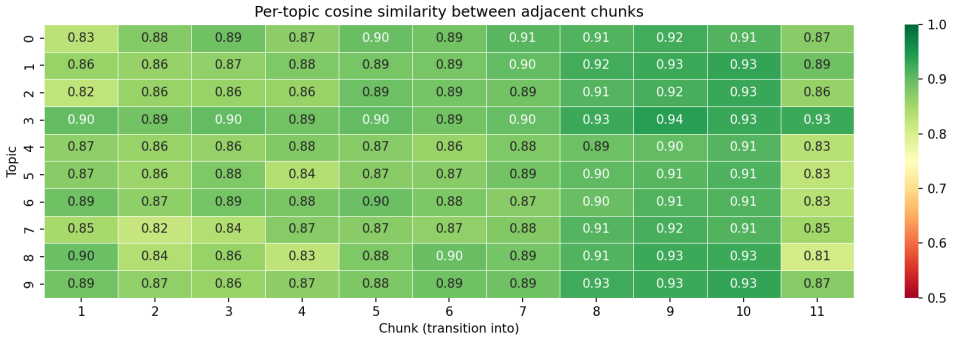


Figure 6: Per-topic cosine similarity between adjacent chunks. The heatmap shows no dips below 0.80 in cosine similarity, indicating stable topics.

stopwords after inference is more transparent and produces similar results as prior removal. Therefore, the relevance metric provides a good balance for words that aren't semantically uninformative but appear across topics.

Relevance was applied at three levels of temporal resolution. For the static topic descriptions, the global topic word distribution ϕ was scored against the full-corpus background marginal $p(w)$. For the per-year descriptions, the same global ϕ was scored against a year-local background, in order to compare terms characteristic of each year with a fixed topic definition. For the per-chunk descriptions, a chunk-specific top-word distribution $\phi_{k,c} \propto \theta_c^\top X_c$ was scored against the chunk-local background, so that each chunk is described in its own terms.

A.1 Domain Stop Nouns

The list below contains uninformative nouns from the corpus in order to prevent them from dominating the extracted topics. Terms are grouped by the rationale for their exclusion.

Designer references: brand, label, house, designer

Vague nouns: idea, thing, way, kind, version, sort, piece, mix, sense, touch, world, view, theme, trend, element, detail, feature, aspect, factor, point, quality, effect, concept, inspiration, influence, mood, vibe, ambiance, aesthetic, atmosphere, impression, good, bad

Runway vocabulary: collection, season, show, look, fashion, clothe, color, fabric, runway, model, style, top, bottom, silhouette, shape, length, pattern, embellishment, dress, garment, clothing, outfit, apparel, wear

Size & quantity: little, long, wide, slim, lot, plenty, bit

Person indicators: woman, girl, man

Temporal nouns: time, day, week, month, year, afternoon, morning, evening, night, today, tomorrow, yesterday

Temporal adjectives: new, old, recent, current, former, previous, early, late, earlier, later, upcoming, first, second, third, final, initial, last, next

Appendix B Appendix B: Topic Modeling Codebook and LLM Provenance

B.1 LLM Provenance Sample Annotation

The subset of reviews was annotated using Chat GPT (`GPT-5.6 Sol Medium`). A temporary chat window was used to prevent any external memory from leaking into the annotation process. The following prompt was used, with the codebook and the sample CSV as file inputs:

"You are classifying fashion runway reviews from the uploaded CSV into one of ten predefined topics (0–9) using the uploaded codebook. Rules:

- Use only the review text as evidence. Ignore any knowledge of designers, brands, or fashion history beyond what the review states.
- The uploaded CSV has two columns: review_id and review. Each row is one complete review; review text is quoted and may contain commas, quotation marks, and line breaks, so parse it as CSV rather than splitting on lines or commas.
- Read a review's full content before deciding. Do not decide from isolated keywords.
- Each topic's exclusion criteria name the topic(s) it is most easily confused with. When a review could plausibly fit two topics, resolve the tie by checking those exclusion criteria directly against the review text.
- Assign exactly one topic to every review. Always choose the single best fit. Assign only a number ranging from 0-9.
- Treat each review independently. Do not let earlier assignments in the batch influence later ones, and do not try to balance the distribution of labels.
- If the batch contains N reviews, return exactly N rows: every review_id from the input must appear exactly once, including ones you find ambiguous or hard to classify.

Output format: Return a plain CSV with two columns review_id, topic_number. One row per input review, in the same order as input. No header commentary, no explanation, no markdown code fences, no extra columns. Output nothing except the CSV."

B.2 LLM Provenance Corpus Annotation

The full corpus was split into batches of 200 reviews, which were annotated using Chat GPT (`GPT-5.6 Sol Medium`). A project folder was created with no access to external memory. The codebook was uploaded as project memory. For each batch, a separate chat was opened. The same prompt as above was used in those chats.

B.3 RollingLDA Topic Codebook

B.3.1 Coding Instructions. Each fashion runway review is classified into exactly one of ten predefined topics, numbered 0–9, using the accompanying codebook. The decision rests only on what the review itself says, setting aside any external knowledge of particular designers, brands, or fashion history not stated in the review. When a review appears to fit two topics, the exclusion criteria in the codebook are consulted: each topic lists the topics it is most easily confused with, and these criteria are checked directly against the review text to break the tie.

B.3.2 Topic 0: Masculine-inflected tailoring and leather. **Description:** This topic covers reviews centered on structured, tailored garments, such as blazers, trousers, and sharp-shouldered jackets, described with an emphasis on masculine, androgynous, or gender-crossing elements. The tailoring is the organizing subject of the review. If leather is mentioned, it functions as a tailoring material (e.g., a leather jacket or trouser) rather than as an accessory or outerwear material.

Inclusion criteria:

Reviews where the primary subject is structured tailoring (blazers, trousers, sharp shoulders), including cases where the structure is conveyed through language like "strong shoulders," or "sharp lines" rather than a garment list; Reviews where the collection or pieces are described using masculine, androgynous, or gender-blurring language, including casual phrasing such as "boyfriend", not only explicit words like "masculine"; Reviews where leather appears as part of the garments as a defining material.

Exclusion criteria:

Reviews centering on fur or luxury outerwear rather than tailored separates, even if some structured pieces are mentioned (see Topic 3); Leather appears mainly in an accessories context (bags, belts), or a single garment is described as leather with no repetition and no structured-silhouette language around it; Reviews where masculine or androgynous language appears early on but the review's own closing assessment is about everyday, practical wearability rather than the gender-blurring point (see Topic 9).

B.3.3 Topic 1: Glamour, eveningwear, and occasion dressing. **Description:** This topic covers reviews of formal eveningwear, such as gowns and cocktail dresses, discussed how they are constructed: silhouette (column, A-line), neckline, beading, draping, and fabric (silk, chiffon, satin). The review focuses on the garment, specifically how it is built and how it moves.

Inclusion criteria:

Reviews focused on gown or cocktail dress construction (silhouette, neckline, beading, draping); Reviews describing formal eveningwear primarily through fabrication and construction detail; Reviews where the bulk of the description is devoted to the garments that fit the previous criteria; Reviews of otherwise mixed-category collections that single out gowns or cocktail dresses as the strongest pieces described through construction detail, even if sportswear, denim, or separates dominate the earlier part of the review.

Exclusion criteria:

Reviews where red-carpet spectacle and the glamour of the event rather than the clothing carries the bulk of the review's attention (see Topic 5); Reviews where lace, corsetry, bra-tops, or a lingerie-derived sensuality is the main focus of the eveningwear rather than gowns (see Topic 6); Reviews of career-retrospective or anniversary shows where the gowns are discussed as artifacts of a designer's history rather than as the focus of the current show (see Topic 4).

B.3.4 Topic 2: Graphic print and nightlife dressing. **Description:** This topic covers reviews of print- and pattern-driven casual clothing, along with disco, party, or nightlife-coded garments, or reviews where pattern and graphic print are explicitly named as the collection's main visual subject regardless of mood. The print or graphic pattern, is where the review's attention lies.

Inclusion criteria:

Reviews built around a print- or pattern-led casual garment or collection (denim, jean pieces, graphic prints, novelty patterns); Reviews describing a disco, party, retro-decade, or nightlife-coded aesthetic; Reviews where pattern and graphic print, rather than cut or fabrication, is the main visual subject even without a disco/party/nightlife addition.

Exclusion criteria:

Reviews fitting any other categories where print makes an incidental appearance; Reviews where casualwear centers more around youth, sport, or athleisure-specific narrative (tracksuits, activewear, logos) rather than a print-led nightlife aesthetic (see Topic 7); Reviews where casualwear centers more around utlity and urban aesthetics (see Topic 9); Reviews built around a soft, pale, floral print rather than a bold, graphic, party-coded one (see Topic 8); Reviews where bra-tops, lace, or corsetry are also named, even just once or twice, alongside the print (see Topic 6) as a specific lingerie-derived garment is a stronger signal than a generic print description.

B.3.5 Topic 3: Luxury fabrics, furs, and outerwear. **Description:** This topic covers reviews of seasonal outerwear, such as coats, made of heavier, luxury materials such as fur, shearling, and mink. The topic applies whenever this vocabulary is the most emphasized in the review, or if garments fitting this description are flagged by the review itself as the standout or best piece.

Inclusion criteria:

Reviews focused on a fur or outerwear garment (coat, shearling, mink, sheepskin); Reviews where a luxury fabric (cashmere, shearling, mink, camel hair, tweed, tartan) defines the collection's identity; Reviews describing seasonal Fall/Winter outerwear as a category in its own right; Reviews of mixed-theme Fall collections (staged, print-led, or concept-driven) where fur, shearling, or coats are named repeatedly, or singled out by the review as the strongest pieces.

Exclusion criteria:

Reviews where leather appears as a tailoring material within a structured, gender-inflected silhouette, and it is the silhouette itself, not the fur or leather as a material, that the review dwells on (see Topic 0); Reviews where a coat or jacket is discussed only for its practical, everyday layering function, with no luxury-fabric language and no repetition (see Topic 9); Reviews where fur or outerwear is mentioned exactly once in passing, within a collection whose main subject is something else.

B.3.6 Topic 4: Designer vision, technical mastery, and intellectual authority. **Description:** This topic covers reviews that evaluate a designer's conceptual, intellectual, or technical skills and ambitions, rather than describing individual garments in detail. The review reads as criticism of the designer's ideas and execution as a body of work, often invoking art history, architecture, or the designer's broader career trajectory. This topic applies only when the designer is the review's sole focus. If the collection's garments are described in more detail than the designer elsewhere in the review (specific fabrics, silhouettes, decorative details), classify by that garment vocabulary instead, even if the review

opens with or repeatedly returns to the designer's stated inspiration.

Inclusion criteria:

Reviews centered on critical evaluation of a designer's concept, artistic or architectural references, or intellectual ambition, where no other topic's garment vocabulary (fur/coat, gown construction, lace/bra-tops, tailoring, print, sport/denim, pastel/pretty mood, separates) is present in force; Reviews that assess a designer's technical skill, innovation, or authority as the main point of the piece (e.g., fabric science, garment construction technique, sustainability process) rather than the garments' category; Reviews discussing a designer's career-long vision or retrospective body of work (e.g. an anniversary), where garments are explicitly framed as artifacts of that career narrative rather than described in their own right.

Exclusion criteria:

Reviews describing the cut and construction of specific garments rather than the designer's conceptual stance (see Topic 0); Reviews where a show's staging, venue, or performance elements, rather than the designer's ideas, are what the review dwells on (see Topic 5); Reviews that name a designer's vision or inspiration in the opening lines but otherwise focus on describing individual garments matching another topic's vocabulary, classify by that garment vocabulary instead; Reviews organized around a single named seasonal concept or theme-title for this particular show, as opposed to the designer's career more broadly (see Topic 5).

B.3.7 Topic 5: Conceptual staging and theatrical collection. **Description:** This topic covers reviews built around a show's staging, performance, or conceptual premise. This may include an unusual venue, live performers, a theatrical set piece, or a single named thematic idea governing the whole show, as the primary thing the critic responds to.

Inclusion criteria:

Reviews centered on the staging, presentation format, or performance concept of a show (venue, set design, live performance elements, props); Reviews where a show's spectacle or premise dominates the review, with garments described mainly in service of that concept; Reviews of shows explicitly framed as performance art, theater, or an immersive experience; Reviews organized around a single named seasonal theme or title (e.g., a one-word concept, an archetype, a titled show), even without a physical set, as long as no other topic's garment vocabulary dominates instead.

Exclusion criteria:

Reviews where formal eveningwear construction, rather than the staging or performance, is what the review dwells on, even if a dramatic setting is named (see Topic 1); Reviews that evaluate the designer's intellectual or artistic vision as a career-long or retrospective body of work, independent of a specific staged event for this season (see Topic 4); Reviews where garment characteristics, are emphasized more independently of the concept of the show.

B.3.8 Topic 6: Lace, lingerie, and sensuality. **Description:** This topic covers reviews built around a lace- and lingerie-derived decorative vocabulary, such as corsetry, bra-tops, ribbon detailing, satin, and sheer fabrication, marking a sensual aesthetic, or reviews invoking mythic or goddess-coded sensuality. The review's attention centers on the erotic, intimate, or mythic charge of the collection.

Inclusion criteria:

Reviews centered on lace, lingerie-inspired construction, corsetry, or bra-tops as the defining characteristics of the garments, even if mentioned only once or twice; Reviews describing a sheer, sensual, or decoratively intimate register (bra tops, slip dresses, exposed boning); Reviews invoking a mythic, "goddess," or old-Hollywood sensuality, whether or not lace or lingerie detailing is also present.

Exclusion criteria:

Reviews centered on formal gown construction (silhouette, neckline, beading) where lace or lingerie detailing is absent or incidental (see Topic 1); Reviews where the register is described only in soft, pale, or romantic mood words ("boudoir," "soft," "romance") with no actual lace, corset, or bra-top garment named: a mood word is not the same as a garment (see Topic 8).

B.3.9 Topic 7: Sportswear, denim, and youth casualwear. **Description:** This topic covers reviews of youth-coded, sport- or street-inflected dressing, such as denim, tracksuits, or activewear, where an association with youth culture, sport, or street style defines the collection. The tone tends to be upbeat and casual, often invoking a specific youth subculture (beach, skate, gym, festival).

Inclusion criteria:

Reviews centered on youthful, sporty, or street-casual dressing (denim, jean pieces, sneakers, sweatshirts, tracksuits); Reviews where tracksuits, activewear, or athleisure pieces define the collection's identity, or where such pieces are introduced as a notable "first" or departure for the brand; Reviews explicitly invoking a youth subculture or age-coded casual setting (teen, beach, skate, festival).

Exclusion criteria:

Reviews built around print- or pattern-led nightlife/disco dressing without a youth-sport or street angle (see Topic 2); Reviews built around practical, separates-led wearability aimed at a mature or professional wardrobe rather than youth culture (see Topic 9); Reviews where a single celebrity or news event, rather than a genuine youth/sport aesthetic, happens to dominate an otherwise unrelated show;

B.3.10 Topic 8: Feminine spring lightness and romantic delicacy. **Description:** This topic covers reviews of a pale, floral, or generally light and pretty feminine aesthetic, described through soft, delicate, whimsical, or romantic language, such as broderie anglaise, crochet, sundresses, and pastel palettes. The mood is light and delicate rather than sensual, formal, or graphic, often invoking innocence, prettiness, whimsy, or a "girl" archetype.

Inclusion criteria:

Reviews centered on soft, pale, or floral dressing (sundresses, broderie anglaise, crochet, eyelet); Reviews describing a romantic, delicate, whimsical, or innocent "pretty" aesthetic, including Fall collections if the palette or mood is explicitly pastel, light, or dreamy; Reviews explicitly invoking a pale color palette (dusty, sherbet, sugar-almond, powder) or lightness as the collection's organizing aesthetic anchor point; Reviews using "soft," "romance," or "boudoir"-adjacent language, provided no actual lace, corset, or bra-top garment is also present.

Exclusion criteria:

Reviews where lingerie and sensuality dominate the review over mentions of softness or romance (see Topic 6); Reviews centered on formal gown construction (beading, column silhouettes, see Topic 1);

B.3.11 Topic 9: Utilitarian separates and wearability. **Description:** This topic covers reviews of separates-based dressing, such as jackets, pants, skirts, and coats, discussed through practical, utilitarian, and urban language. The review's attention centers on the separate pieces themselves and their everyday wearability, including cases where the review's own conclusion frames the collection as practical, wearable, or "for a well-dressed woman to throw on," even if other vocabulary (androgynous framing, leather) appears earlier.

Inclusion criteria:

Reviews centered on separates-based dressing (jacket, pant, skirt, coat) as the main subject; Reviews describing utilitarian or urban language applied to separates; Reviews framing a collection around wearability or practicality rather than spectacle, luxury status, or concepts; Reviews emphasizing practicality and everyday dressing.

Exclusion criteria:

Reviews where fur or seasonal luxury outerwear is named repeatedly or flagged as the standout piece, rather than a coat being discussed function (see Topic 3); Reviews where tailoring carries an explicit masculine or gender-blurring charge as the review's sustained point, not just an incidental phrase (see Topic 0); Reviews where separates are styled specifically for a youth or street-casual audience, or introduce denim/sneakers as a notable choice, rather than general practical wearability (see Topic 7).

Appendix C Appendix C: Event Cluster Interpretation and Mapping

C.1 Event Cluster Labeling

Event cluster labeling was done across three levels of abstraction. At the finest level, the 30 clusters were assigned descriptive subtype labels by inspecting each cluster's top verb lemmas by frequency, and naming the semantic pattern they expressed. These 30 subtypes were then mapped onto two independent higher-order schemes. The first, an agency dimension, tagged each subtype according to whose action the verbs encode, distinguishing events attributable to the Designer/ Collection from those belonging to the Critic, with a residual ambiguous category reserved for corpus-dominant clusters with no clear semantic signal. (Being & Presentation, Generic Action). The second scheme applied Motta-Roth's [36] rhetorical move taxonomy of four categories: introducing, outlining, highlighting, and evaluating.

C.2 Event Cluster Mapping

Table 6: Event topic labels and their annotation scheme.

#	Label	Top verbs	Motta-Roth	Agency	Group
1	Design & Ornamentation	design, draw, layer, decorate, print, embroider, accessorize, paint, display, render	Outlining	Designer/Collection	Craft
2	Desire & Affinity	love, like, embrace, lay, touch, marry, seek, attract, crave, hug	Evaluating	Critic	Stance
3	Rejection & Departure	leave, throw, abandon, explode, fade, escape, disappear, kill, shed, ditch	Highlighting	Designer/Collection	Contextual Positioning
4	Being & Presentation	be, show, inspire, prove, provide, talk, expect, let, reflect, speak	Outlining	Ambiguous	–
5	Visual Effect	pop, snap, flare, peek, shine, flash, fuse, blossom, glow, spark	Outlining	Designer/Collection	Appearance & Movement
6	Allusion	miss, define, nod, found, lighten, hew, hark, wove, liven, hail	Introducing	Critic	Observation
7	Evocation	emerge, evoke, emphasize, rise, conjure, evolve, embellish, inject, elevate, embody	Highlighting	Designer/Collection	Contextual Positioning
8	Generic Action	have, take, want, get, give, use, put, open, add, call	Outlining	Ambiguous	–
9	Spatial Placement	move, sit, cover, represent, debut, march, land, place, visit, parade	Outlining	Designer/Collection	Appearance & Movement
10	Continuation & Development	work, keep, continue, produce, manage, live, focus, remain, grow, develop	Introducing	Critic	Stance
11	Placement	turn, stand, fall, push, stick, break, hang, shape, float, range	Outlining	Designer/Collection	Appearance & Movement
12	Reinterpretation	return, recall, replace, revisit, reappear, repeat, reinterpret, reprise, rework, re	Highlighting	Designer/Collection	Contextual Positioning
13	Resistance	appeal, avoid, resist, distract, deny, detract, hide, ignore, refuse, eschew	Evaluating	Critic	Stance
14	Initiation	come, go, include, start, become, appear, bring, begin, suggest, create	Highlighting	Designer/Collection	Contextual Positioning
15	Direction & Showcase	build, launch, lead, capture, top, aim, base, express, showcase, model	Evaluating	Critic	Stance
16	Cognition	say, do, know, explain, think, mean, end, try, happen, imagine	Introducing	Critic	Observation
17	Tailoring	cut, wrap, line, tuck, edge, tap, trim, head, whip, nail	Outlining	Designer/Collection	Craft
18	Fluid Motion	fill, flow, wash, destine, filter, spill, delve, pour, wind, dispense	Outlining	Designer/Collection	Appearance & Movement
19	Construction & Composition	make, translate, construct, underscore, yield, spell, precede, complement, juxtapose, derive	Outlining	Designer/Collection	Craft
20	Garment & Styling	wear, dress, drape, style, knit, weave, suit, sew, tailor, lace	Outlining	Designer/Collection	Craft
21	Competition	play, strike, lose, win, match, kick, sport, settle, dominate, toss	Evaluating	Critic	Stance
22	Perception	look, seem, feel, see, find, describe, read, tell, decide, ask	Outlining	Critic	Observation
23	Refinement	veer, patchworke, hone, ampe, soundtracke, imbue, vie, tout, devise, evince	Outlining	Designer/Collection	Craft
24	Ornamentation	mark, catch, bear, dub, stray, toy, stud, garner, meld, billow	Outlining	Designer/Collection	Craft
25	Traversal Motion	walk, run, explore, step, reach, fly, slip, drive, trail, travel	Outlining	Designer/Collection	Appearance & Movement
26	Sonic Metaphors	sound, echo, hear, riff, sing, scream, resonate, listen, accent, pronounce	Introducing	Designer/Collection	Contextual Positioning
27	Transformation	change, extend, transform, expand, soften, morph, refine, scale, tweak, sharpen	Highlighting	Designer/Collection	Contextual Positioning
28	Combination	mix, combine, gather, separate, pack, merge, zip, divide, split, pile	Outlining	Designer/Collection	Craft
29	Runway Motion	stag, crop, sweep, strut, stomp, dig, pluck, finesse, scoop, skim	Outlining	Designer/Collection	Appearance & Movement
30	Naming & Noting	feature, present, reveal, reference, lack, form, name, notice, note, hint	Introducing	Critic	Observation

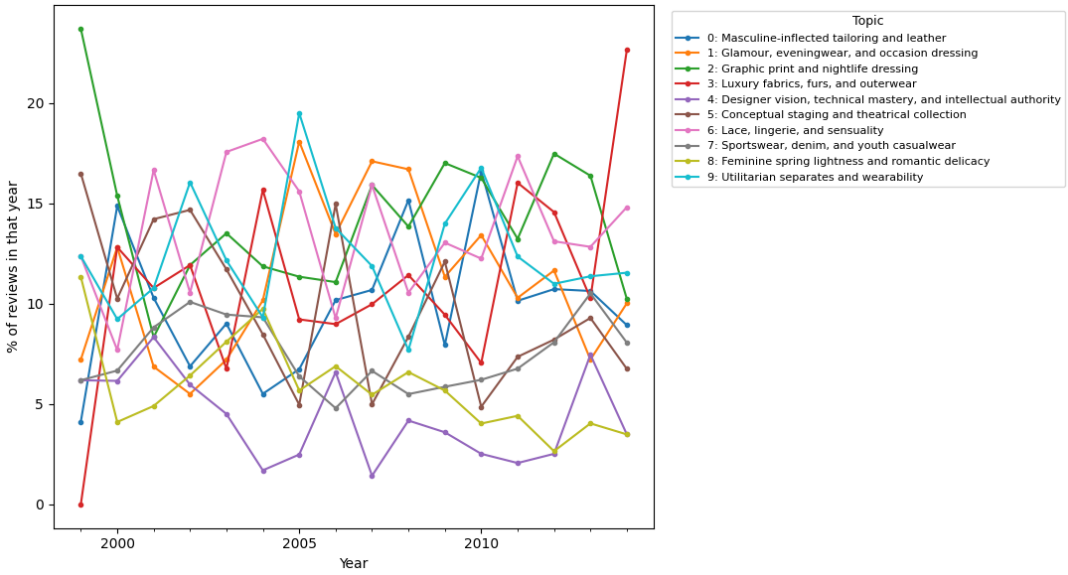


Figure 7: Topic share per year shown as a combined overview.

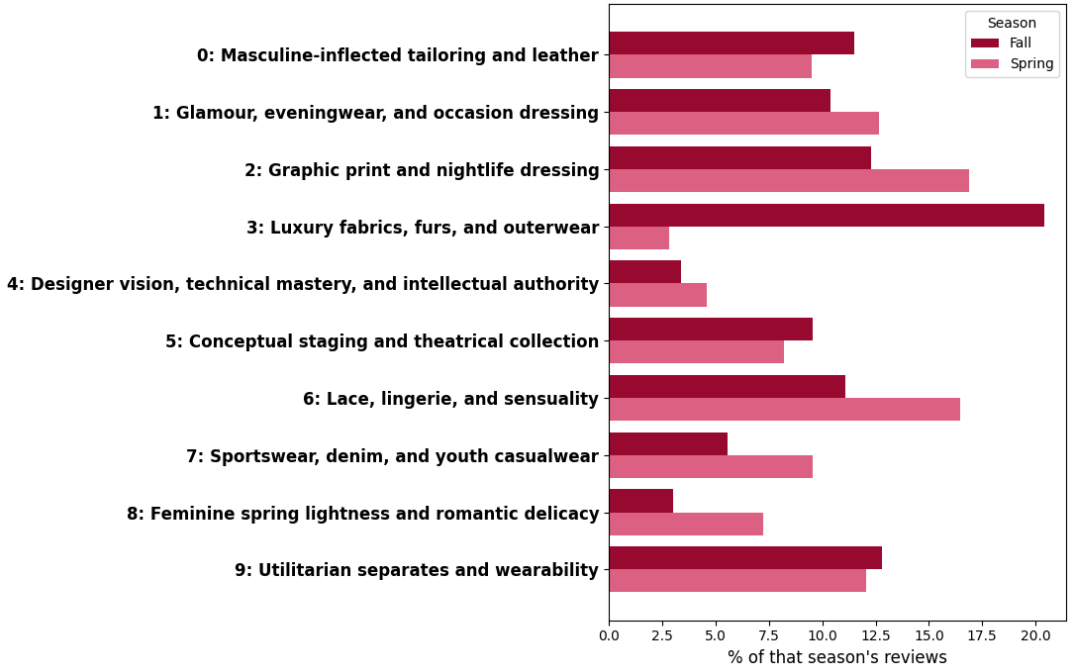


Figure 8: Topic share per season, shown per topic.

Appendix D Appendix D: Supplementary Materials

Appendix E Appendix E: Generative AI Statement

Generative AI (GenAI) was used as a tool during this thesis for both coding, writing, and annotation tasks.

E.1 Coding

In the development of code, I made use of GenAI for the following tasks: translating formulas from literature into working implementations, resolving bugs, and answering syntax questions. Where no existing library was suitable, it also assisted in adapting a method from its formal definition into python, the PERMANOVA test is for instance implemented from first principles with and eta-squared statistic scored against a group-label permutation null and a matching dispersion check, rather than drawn from an existing package. It also provided support in implementing the RLDA framework to the corpus, which needed to be converted from its implementation in R. It was also used to ask for corrections on syntax for plots, for example when a specific implementation did not match up with the expected result. In each case outputs were verified against intended behavior: the statistical procedures against their formal definitions and, where possible, against reference implementations. Lastly, it was used to generate LaTeX format tables from results, in which case outputs were also verified to ensure no hallucination took place.

E.2 Writing

GenAI was also used to aid in writing, by checking spelling and grammar and to generate feedback on clarity and structure. The arguments, interpretations, and conclusions are entirely my own.

E.3 Annotation

GenAI was used for annotation of labels, the full procedure of which is described in Section 3.