Readers Negotiating Genre: Semantic Space in Children's Literature

Yauheniya Lekarevich, Sergei Pashakhin

The Lion and the Unicorn, Volume 48, Number 2, April 2024, pp. 165-184 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/uni.2024.a980317

Readers Negotiating Genre: Semantic Space in Children's Literature

Yauheniya Lekarevich and Sergei Pashakhin

The inquiry into the nature of literary genre is vast and elusive, lacking definitive answers. As far back as 1938, the Russian philologist Boris Yarkho astutely observed the absence of a systematic approach to genre study, affecting both genre definitions and classifications (50). This observation remains relevant today, as the understanding of literary genres continues to present challenges. In this article, we explore children's literature genres using computational methods to assess their potential in modeling theoretical genre concepts while acknowledging the absence of definitive resolutions.

James Gifford argues that we gain more from fluid definitions than we lose from rigid categorization. Leiderman notes that during the modern period, genre's significance declined as the traditional system of canonical genres transformed. He highlights the oscillating nature of genre concepts, preoccupying scholars in the 1970s but falling to the sidelines in the post-structural period of the 1980s (Leiderman 10–11). Still, genre criticism has experienced a revival in recent years, partly due to computational text analysis methods (Underwood, "Genre Theory" 3).

Scholars have debated the perception of children's literature as a genre. For instance, Perry Nodelman treats it as a genre in his book *The Hidden Adult*, but Kimberly Reynolds (210) and Marah Gubar (214) challenge this notion. Regardless, researchers effectively utilize various categories such as picture books, children's poetry, fantasy, science fiction, and graphic novels to group and analyze literary texts.

In this study, we concur with Reynolds's view and emphasize the importance of examining genres within children's literature. Recognizing genres' historical and dynamic nature, we aim to explore the resemblances among realism, fairy tales, sci-fi, and fantasy in Soviet and post-Soviet children's literature. Due to the lack of a consensus in scholarly categorizations, we turn to readers' genre classifications for our investigation.

The Lion and the Unicorn 48 (2024) 165–184 © 2026 by Johns Hopkins University Press

History of Genre Studies

The study of genre traces its origins back to classical antiquity, notably credited to Aristotle. Normative poetics in the nineteenth century imposed strict boundaries on genres through regulations and artistic rules. The Romantic period introduced profound philological theories, reshaping literary theory. Even in premodern times, empirical genre studies attempted to compile taxonomies (Frow 58). Canonical literary forms gradually started to coexist with noncanonical ones like novels and romantic poems, deviating from observable and describable structural bases (Tamarchenko 11). Darwin's theory contributed a biological metaphor for genre structure; the metaphor, largely abandoned during the greater part of the twentieth century, is now resurfacing in the mainstream under the notion of "cultural evolution theory" (Sobchuk and Šeļa, "(Distant) Reading").

The late nineteenth century witnessed a shift from normative genre theory to empirically oriented research, emphasizing genre's historicity. The Russian Formalist school, emerging in the early twentieth century, inherited the concept of literary evolution, highlighting the perpetual mutability of genre hierarchy and the significance of both form and function (Duff 7). While influential, the school faced criticism from contemporaries like Olga Freidenberg, who questioned its formalism and Darwinian influences (11). Freidenberg explored the connection between morphology and semantics, suggesting that genre classification is arbitrary due to its intimate link with the plot (13), echoing the formalists' ideas on genre fluidity.

Methodologically aligned with formalism, the work of Leningrad scholar Vladimir Propp on the morphology of fairy tales laid the foundation for subsequent structuralist research. Beyond Leningrad, Polish formalism represented by Ireneusz Opacki and Czech structuralism, as well as Tartu's structural semiotics, spread influential ideas, although they shifted focus away from genre. In the Anglophone world, Canadian critic Northrop Frye developed a distinct kind of structuralism, exploring the correlation between mythology and literature (Leiderman 74). The Konstanz School in Germany, led by Robert Jauss, redirected attention to reader expectations shaped by genres. Mikhail Bakhtin contributed significantly to genre theory through his work on the novel and the concept of "genre memory." Tzvetan Todorov's attention to Bakhtin's works influenced French structuralists and remains influential to this day, reaching scholars such as Franco Moretti, John Frow, Tony Bennett, and Fredric Jameson. Todorov's own theory of genre blurred the distinction between formalism and structuralism, particularly in the area of genre theory (Duff 11).

Naum Leiderman presents a comprehensive "genre disposition" that characterizes the modern field of literary study, classifying theorists into four distinct groups (12–14). The normative theory, advocated by scholars like Jean-Marie Schaeffer, promotes a "taxonometric concept." Relativistic concepts, ranging from Jacques Derrida to Alastair Fowler, assert the constant changeability of genre as a fundamental law. The third group delves into structural and semantic genre problematics, with scholars like Stefania Skwarczyńska, Tzvetan Todorov, Aviva Freedman, and Peter Medway. Lastly, the genetic direction focuses on establishing the semantics of genre forms, exemplified by scholars like Northrop Frye, Paul Hernadi, and Mikhail Bakhtin.

Genre exploration sparks captivating discussions of considerable complexity and extends into various fields such as films and music. Given the wealth of theoretical research in general literature studies, it is intriguing to examine the distinctions within the field of children's literature.

Genres in Children's Literature Studies

The *International Companion Encyclopedia of Children's Literature*, edited by Peter Hunt, offers a compelling repertoire of children's genres. This volume explores the origins and modern state of various genres, including poetry and drama, under the section "Types and Genres." Reynolds, in her book, endeavors to capture the inherent specificities of the children's literature genre system. She posits that genres migrating from adults' literature to children's literature are not merely preserved but restored and rejuvenated (85). Tri Pramesti's proposition of a taxonomy for children's and young adult literature subgenres represents an uncommon attempt in the field, where genre primarily serves to structure the material (8).

In Russian children's literature, Olga Oktyabrskaya's dissertation on the genre system in Soviet literature from the 1920s to the 1950s is a significant contribution. In both Russian and Anglo-American scholarship, the structuralist school wields significant influence. Peter Hunt notes that "structuralism, although somewhat outdated as a critical trend, may serve as a starting point for the study of myth, legend, folk and fairy tales—but only a starting point" (Understanding 9). Children's literature studies borrow methodologies from the broader field of literature, such as Maria Nikolajeva adapting Lotman's and Todorov's findings. However, there also might be a sense of isolationism within the field, despite the ongoing advancements in quantitative genre studies.

Digital Approaches in Genre Studies

The digital research on genre-related issues is relatively recent, with the pioneering work emerging as late as 1988 (Calvo Tello 54). During the late 1990s, the field began to take shape, and in the latter half of the 2000s, it witnessed the emergence of numerous influential publications on the topic. Although methodologies from digital humanities have been applied to children's literature, resulting in captivating outcomes (Adukia et al.; Maslinsky), computational methods for genre studies in children's literature are still relatively underexplored, except for Dunst and Hartel's notable study on the visual styles of different genres in the related field of comics.

There are several approaches to the genre classification task. Research often relies on lexical methods, where the most frequent or distinctive words are utilized to differentiate genres, as demonstrated in Ted Underwood's research (Underwood, "Life Cycles"). Another approach involves topic modeling, enabling researchers to delineate the thematic domains of the text and use topics as distinctive features for classification, with examples of this approach evident in the work of Christof Schöch. A notable combination of these two is proposed by Kirill Maslinsky in his case study of animals in children's literature, where a vocabulary growth model is applied to infer how vocabularies vary with genre (Maslinsky "How Exactly Does Literary Content Depend on Genre?").

A different route is network analysis, a method that automatically extracts links (defined in various ways, such as shared dialogues or spatial connections within scenes) to detect communication patterns among characters in specific texts. Rahul et al. propose such work, where a corpus of fanfiction is utilized to build character interaction graphs to train a classifier to predict genre. It is also not uncommon to blend methods to enhance analysis. Lena Hettinger et al., in their article dedicated to German novels, employ three approaches to test the accuracy of the classifier, ultimately favoring a thematic approach to character networks and also using stylometric features. In stylometry, both lexical and non-lexical features can be employed, encompassing word frequencies, punctuation usage, sentence length, and more.

Another strategy to utilize lexical analysis for genre categorization involves the examination of distinctive words. The notion behind this approach lies in the existence of words that are exceptionally characteristic of a given collection, and their distinctiveness can be measured. Keli Du and her colleagues have recently adopted this analytical path in their article, which tackles the task of novel classification. They explore nine different corpora in seven languages, utilizing frequency, distribution, and dispersion-based measures to identify distinctiveness.

The aforementioned approaches primarily rely on classification techniques, although the utilization of clustering appears relatively rare. While classification assigns data points to predefined classes based on labeled training data, clustering searches for groups (clusters) in data without prior knowledge about classes. An attempt at clustering was made by Mariona Coll Ardanuy and Caroline Sporleder, where they sought to organize texts by genre using character networks.

While most computational methods employ lexical approaches as the foundation for genre comparison, alternative possibilities also exist (as exemplified by social network analysis and partially by stylometry). A semantic approach extends beyond lexical analysis by examining not just individual words but also the relationships and conceptual frameworks within the text. Although not extensively prevalent in genre studies, word embedding techniques have been applied to nonliterary data (Sethy et al.). For example, Eran Ozsarfati et al. use embeddings on book titles rather than content and report moderate success in classification.

A rather promising approach is demonstrated by Maria Antoniak et al., who investigated the problem of genre classification through the lens of readers' perceptions. They assessed this aspect by utilizing the genre tag systems of *Goodreads*, as well as by conducting another study alongside Melanie Walsh, focusing on reviews. In these works, the literary text itself also was not explored; instead, the research sought to understand the readers' comprehension of the genre system.

In this study, we employ a similar approach by acquiring user-generated data and training the classifier on document embeddings of full text. These design choices appear to be relatively underexplored within the current state of the field.

Data and Methodology

In this work, we adopt an exploratory design that abstains from offering conclusive solutions. Our exploration relies on two primary sources: data scraped from a prominent Russian-language website *FantLab* (an analog of *Goodreads*), which offers user-generated content on literary works, and the *Corpus of Russian Prose for Children and Youth* (*Detcorpus*), containing around three thousand texts compiled by Maslinsky et al.[1]

FantLab contains user-assigned tags for features of literary texts like plot linearity and narrative devices, selected through popular vote. We collected entries with user tags such as "realism," "fairy tale," "sci-fi," and "fantasy" within the categories of "children's literature" and "teenage literature." To avoid unreliable markup, we limit our data collection to the top ten pages of results in each category.[2]

There is a difference between the two datasets: while *FantLab*'s primary focus is fantastic in a broad sense, *Detcorpus* mainly consists of realistic fiction. Originating as a corpus of mid-twentieth-century literature, *Detcorpus* has expanded to include texts from 1900 to the 2010s across various genres, though some, like "fantasy," remain underrepresented. On the other hand, *FantLab* generally represents readers' literary demands but does not contain, for instance, modern romantic fiction for girls. In our data, the overlap between datasets is 443 texts, on which researchers' corpus-building decisions and readers' genre perceptions intersect. *Detcorpus* contains internal genre categories, but they may lack a common and reliable basis, making *FantLab*'s popular vote procedure for genre attribution seem more valid.

First, we examine reader-perceived genres on the level of user-assigned tags. Tzvetan Todorov, drawing upon prior research, articulated several crucial aspects that shape our methodology. Firstly, he posited that genres exist as institutions, acting as "horizons of expectation" for readers and "models of writing" for authors. He underscored the necessity of perceiving genres historically, emphasizing their temporality, a concept he termed "codification" (Duff 198). Within this statement, two pivotal notions emerge: the historical context of genre and its function. Thus, we opt for mapping genres in the space of user-assigned features (tags) with t-SNE—a technique aiding visualization of complex data—to provisionally confirm that the data have a signal that might help to learn patterns of genre assignment. Then, we use machine learning algorithms to compare the explanatory value of user-assigned tags capable of shaping the genre signal. We estimate tag importance with three metrics (Chi-square, F-values, information gain) to test measurement reliability and to learn effect sizes (the magnitude of feature-genre association).

Then, we go to a level below user-assigned tags and investigate genres in relation to the content of texts. Todorov further employed an operationalization of the genre based on "discursive properties," encompassing the semantic, syntactic, and verbal aspects of the text. While somewhat broad in scope, this operationalization provides a starting point for implementing theory into computational practice. Within this article, we use three techniques: first, to capture the semantic dimension, we employ doc2vec embeddings (Le and Mikolov). Second, to represent the verbal (lexical) dimension, we use topic modeling with latent Dirichlet allocations (LDA) (Blei et al.). Third, to represent the syntactic level, we use stylometric attributes of individual texts. We elaborate on each technique below.

We build upon experiments reported by Sobchuk and Šeļa and rely on the best-reported combinations for text preprocessing and data representations ("Computational Thematics" 9); we also benefited from their code for document embedding. As reported, the best performance is provided by "strong

thematic foregrounding" combined with doc2vec embeddings or LDA with one hundred topics built for the five thousand most frequent words. Strong foregrounding consists of lemmatization and removal of proper names and all words that are not nouns, verbs, adjectives, or adverbs ("Computational Thematics" 5). Originally, strong foregrounding included lexical simplification, but we had to omit it from our study due to the lack of high-quality language resources for Russian.

Unlike the syntactic way of representing a document, the semantic and the lexical ways rely on more complicated computational tools. Doc2vec is a way to represent a corpus with numeric vectors for a text. It is designed to capture document similarity in the space of a language model. A language model aims to represent the relationships between individual words. Following Sobchuk and Šeļa, we used a precomputed *FastText* model for the Russian language to find document embeddings. A document is represented with a vector of three hundred numbers computed, based on the language model and the document content. We used this numeric representation of a text from *Detcorpus* to train classifiers and explore genre structures. Finally, to estimate genre similarity, we group our texts by genre and measure their average similarity by doc2vec embeddings with cosine similarity on both training data and the classified *Detcorpus*. The LDA topic model is similar to doc2vec but less complex. LDA aims to extract a number of topics or themes unknown beforehand. It searches for the most optimal way to put words—as they are used in a corpus—into the predefined number of topics. This search produces two matrices: word-topic and document-topic probabilities. We used the latter to explore the performance of classifiers and to validate genre mapping.

To use a trained classifier, we must make several decisions about algorithm selection and data transformation. A classifier—trained either on semantic, lexical, or stylometric attributes—aims to learn how *FantLab* users assign genre. To locate the primary information source about user decisions, we compare semantic, lexical, and stylometric classifiers. The dataset with 443 matched works has a class imbalance, with only 29 examples of fantasy and 92 of sci-fi (table 1). To address this issue, we limited the pool of algorithms to the simplest that have facilities to weigh minority classes or address the imbalance in other ways. Thus, we compare the performance of logistic regression, random forest, decision tree, and balanced bag of histogram gradient boosting (BBHGB) classifiers against each other and a dummy classifier predicting only the majority class. Additionally, we explore six standard techniques of data preprocessing to overcome the class imbalance: random undersampling, synthetic minority oversampling technique (SMOTE) (Fernandez et al.), borderline SMOTE, SMOTE with SVM sample detection (SVMSMOTE), SMOTE with Tomek link function, and SMOTE with ENN cleaning (SMOTEENN).

Table 1. Genre and narrative forms distributions in the matched Detcorpus texts. *10 texts under the category "cycle" were removed, as they required the grouping of several texts.

Genre	Number of documents	Avg. N words	Avg. sentence length	Avg. sentence length variance
Realism	206	13,807.86	8.57	45.33
Fairy tale	116	11,308.83	8.19	32.91
Sci-fi	92	36,425.84	8.64	40.24
Fantasy	29	69,396.51	8.78	39.86
Total:	443			
Novella	180	35,603.38	8.45	38.67
Short story	154	2,184.05	8.45	48.21
Fairy tale	62	1,250.51	8.77	28.17
Novel	25	95,364.88	9.03	45.05
Microstory	12	875.41	7.86	31.2
*Total:	433			

Additionally, we use the opportunity presented by *FantLab* data to explore user categorization of the narrative forms. Building upon Todorov's observation that the novel is no longer perceived as a distinct genre—he wrote: "no longer quite genre in our eyes"(194)—we examine user-assigned tags in *FantLab*'s "genre and theme classifier" section, which designates narrative forms such as novel, novella, and short story. For the sake of consistency with user-generated content in this study, we refer to these divisions as narrative forms. The affordances of *FantLab* prioritize thematic classification, followed by the classification of form.

As for the segregation of "narrative forms" (novel, novella, short story, etc.), Yuri Tynyanov noted, "We tend to name genres by *secondary resultant attributes*, roughly speaking, by magnitude" (Tynyanov 105). Unsurprisingly, our research echoes this perspective, foregrounding the reader's understanding in our analysis. Nevertheless, this understanding may not consistently align with literary theories, as readers often use "fairy tale" both as a genre tag and a category for narrative form or employ the uncommon term "microstory." We construct the syntactic classifier to address these secondary attributes. For each document in *Detcorpus*, we calculated stylometric attributes, encompassing word count, sentence count, average word length, total number of punctuation symbols and their respective types, average sentence length

with variance, and counts for each part-of-speech category. These data were employed to investigate the narrative forms of texts.

Results

Our strategy for building the genre classifier resulted in a combination of data preprocessing and an algorithm showing good quality (table 2). Typical quality for such algorithms ranges between fifty percent and ninety percent (Antoniak et al.; Hettinger et al.). We followed the standard procedure for training and evaluation: an algorithm takes a portion of data for learning and then tries predicting the other portion. We estimate performance by randomly splitting data into training and testing parts and averaging quality metrics over several trials. For each combination of an algorithm and preprocessing, we did ten such trials. Reported metrics range between 0 and 1, where the value of 1 stands for perfect performance. Precision is the ability to label positive examples as such, and recall is the ability to find all the positive examples. Here F1 is a harmonic mean of precision and recall, aiming to summarize the quality of a classifier.

Table 2. The top performing classifiers and their quality metrics for each class. Metrics are averaged over 10 runs (10-fold cross-validation)

Best classifier	Space	Genre/Form	Precision	Recall	F1	Avg. F1
SVMSMOTE + Logistic regression	Semantic (doc2vec)	Realism	0.88	0.91	0.89	0.82
		Fairy tales	0.90	0.77	0.83	
		Sci-fi	0.83	0.86	0.84	
		Fantasy	0.62	0.85	0.71	
SMOTETomek + BBHGB	Thematic (LDA)	Realism	0.71	0.78	0.74	0.70
		Fairy tales	0.58	0.48	0.52	
		Sci-fi	0.84	0.81	0.82	
		Fantasy	0.71	0.79	0.73	
SVMSMOTE + Random Forest	Stylometric	Novella	0.91	0.94	0.93	0.67
		Short story	0.87	0.84	0.85	
		Fairy tale	0.59	0.59	0.58	
		Novel	0.73	0.75	0.72	
		Microstory	0.37	0.32	0.29	

We observe that the best quality is provided by semantic space, followed by thematic and stylometric attributes. Most importantly, the classifier trained on doc2vec embeddings has more consistency in quality across classes. Thematic and stylometric data representations fail to differentiate between certain classes with quality close to and below simple random assignment (0.5). However, the inability to predict the user-labeled category of "fairy tale" might be due to its inconsistency with labels like novel, novella, and short story. The challenge in recognizing the "microstory" may stem from the very small number of examples available for training (N=12). Overall, semantic information performs more effectively in understanding how *FantLab* users assign genres.

Semantic Space

Semantic space is a way to quantitatively represent works from *Detcorpus*. By reducing rich text data to a number of vectors, we gain the ability to examine how literary works are grouped and how well these groupings match genres, both voted by *FantLab* users and estimated with our classifier. Using t-SNE, we reduce the document vectors to two-dimensional space; the visualization allows us to make the following observations.

Only fantasy and sci-fi demonstrate distinct and well-defined clusters, whereas fairy tales appear to intermingle with realism, defying conventional expectations. Surprisingly, we are unable to identify a stronger resemblance between fairy tales, en masse, with either fantasy or sci-fi, despite the inherent expectation of a "fantastic" element within them. One plausible explanation for this phenomenon could be the limitations of our dataset and the issue of class imbalance. Alternatively, the inconsistencies within the reader-marked-up class itself might play a significant role. For instance, the texts can vary greatly in length, ranging from short narratives describing protagonists as animals (for example, V. Bianki's works) to more intricate examples depicting the magical exploits of human and nonhuman protagonists, as seen in Alexander Volkov's adaptation of *The Wizard of Oz*.

Documents do form clusters based on genre in the semantic space. The most evident visual clusterization is observed on the t-SNE representation of *FantLab* data, where documents are expressed with user-assigned features (figure 1).[3] In figure 2a, the semantic similarity of texts with user-assigned genres reveals two distinct clusters of sci-fi texts. One of these clusters is in proximity to fantasy, while the other is closer to fairy tales and realism. Notably, this second cluster is associated with the works of Kir Bylychev, implying the author's semantic affinity to realistic science fiction, particularly his saga of the cosmic travel of a girl named Alice.

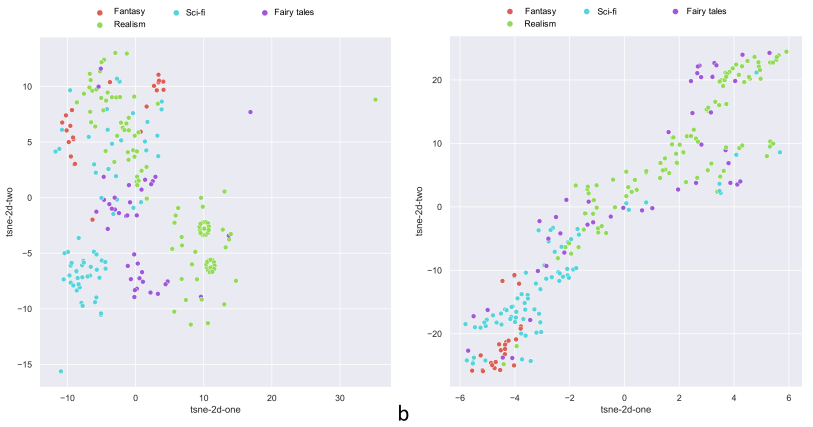

Fig. 1. Genres in the user-generated space: a) with year of publication excluded; b) with year of publication included.

Fig. 2. Genres in doc2vec embeddings: a) the text matched FantLab (N=443); b) Detcorpus with genre classified by the algorithm (N=2827).

In the sci-fi and fantasy cluster, literary fairy tales like Lazar Laginov's *The Old Man Hottabych* and Evgeny Shvart's *Absent-minded Wizard* are situated towards the boundary of realism, while Sof'ya Prokofieva's *The Patchwork and the Cloud*, Vitali Gubarev's *Kingdom of Crooked Mirrors*, and others lean more towards fantasy. This positioning suggests their potential as forerunners of Russian fantasy, while we need to acknowledge the genre's reliance on translated works from the Western tradition (Kovtun). However, our classification of *Detcorpus* using semantic features still results in significant misclassifications. Numerous texts that are unmistakably realistic are mislabeled as fantasy and sci-fi, impeding clear clusters in figure 2b. Despite this limitation, our observations indicate that semantic similarities align with the assigned genre, signifying users' reliance on such associations.

On closer inspection, doc2vec not only groups texts by genre but also by author. For instance, figure 2a displays the clustering of works by modern writers Dmitry Emets and Natalia Shcherba, both immersed in the fantasy genre. Meanwhile, figure 2b consolidates the works of Mikhail Prishivin, Petr Zamojsky, Konstantin Tomashevsky, and S. Ventsel', forming a cluster primarily composed of very short stories from the 1920s to the 1940s, some pertaining to the animalistic genre, albeit misclassified as fantasy.

The impact of the year of publication on the data's structure is striking (figures 1 and 3). Figure 3a–b reveals the shifts in genre paradigms throughout the history of Soviet and post-Soviet children's literature. Moreover, the predominance of temporal signals over textual signals is also noticeable, with the period of publication exerting a powerful influence on the overall layout, overpowering certain subtle semantic features. As previously discussed, the prominent trend in children's and young adults' literature is the transition from realism to a prevalence of mass genres (sci-fi and fantasy among them), while fairy tales sporadically emerge throughout. Figure 3b demonstrates the emergence of distinct clusters within specific periods, such as the distinctive clustering of works in the 1920s and 1950s, but due to the consistent misclassification of genres, it becomes challenging to determine if these divisions truly represent distinct genres. Although the entirety of the period spanning from 2000 to 2017 appears to be clustered together in a cohesive space, it is imperative to recognize that the literary works situated at its boundaries may, in fact, possess considerable semantic disparities.

Genre Similarity, Feature Importance, and the Dramatic Effect of Time

We classified *Detcorpus* and assessed the similarity of genres using doc2vec embeddings (table 3). While the user-labeled data indicates the highest similarity for fantasy and sci-fi, it registers as moderate on the cosine similarity

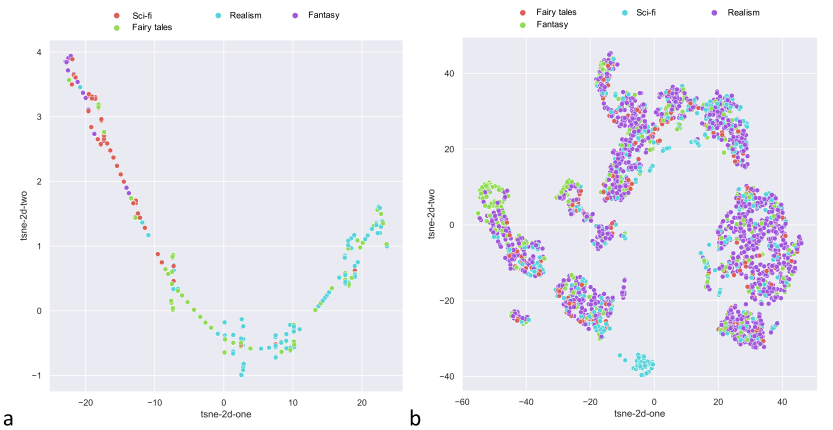

Fig. 3. Genre mapped in doc2vec with year of publication: a) FantLab data; b) Detcorpus classified by the algorithm.

scale. It should be noted that comparing *FantLab* to *Detcorpus* data highlights a significant disparity, revealing the limitations of our classifier. Upon manual examination of texts within these categories, hundreds of errors in automatic genre labeling were discovered. Nevertheless, the measurements conducted on *FantLab* reliably indicate a relatively weak similarity among genres like realism, fairy tales, and sci-fi.

Table 3. Genre similarity measured as average cosine similarity of doc2vec vectors for documents in each genre: a) FantLab matched to Detcorpus (N=443); b) Classified Detcorpus (N=2827).

a)	Realism	Fairy Tales	Sci-Fi
Fairy tales	0.20		
Sci-Fi	0.20	0.21	
Fantasy	0.19	0.21	0.29

b)	Realism	Fairy Tales	Sci-Fi
Fairy tales	0.17		
Sci-Fi	0.11	0.13	
Fantasy	0.14	0.10	0.002

To elucidate patterns of genre assignment, we estimated the effect sizes of user-generated features for predicting each genre. The year of publication stands out as a primary feature ($p < 0.05$, Cohen's $w = 0.7$), with a compelling literary-historical explanation. Fantasy, horror, and other mass literature genres surged after the decline of the Soviet centralized publishing system, once constrained by the ideological apparatus.

Noteworthy features characterizing the texts include the period of action, plot linearity, target audience, and narrative form. Regarding audience and narrative forms, we observe that they are connected and also related to plot types. Smaller forms like novellas and short stories prevail among the children's audience, while novels and literary cycles are more frequent in writing for teenagers. The most pronounced effect pertains to the time of action: realistic texts center on the twentieth century, in fairy tales the label "undefined time" prevails, and fantasy often unfolds in the twenty-first century, less frequently in "Ancient world" and "Late Middle Ages/Renaissance" (figure 4b). Science fiction displays the most temporal diversity, embracing far and near future narratives across all categories, offering a rich tapestry of plots—linear, linear with excursions, linear-parallel, parallel, and nested (decreasing). Similarly, fantasy exhibits diverse plot structures, whereas realism and fairy tales strongly favor linear plots, the most common in all categories (figure 4a). The intricacy of the plot may intertwine with the audience's age, as text simplification might be used to align text with presumed children's development.

Conclusion

Thomas Winner, echoing Yury Tynyanov's insights, asserts that each genre, in a given historical period, forms a system that is synchronic in character. However, it is simultaneously an ever-evolving series, rendering it diachronic (265). This valuable observation guided our study's design, as we aimed to incorporate these temporal dimensions of genres along with textual features.

Notably, the effect size of the variable "year" that we uncovered is remarkably significant, overshadowing other features to a considerable extent. Consequently, the layout of our visualizations depicting relationships between texts underwent dramatic changes upon its inclusion, as we attempted to "combine historical testimony and text in a single model" (Underwood, "Genre Theory" 5).

Moreover, Benjamin Gittel's landmark study provides evidence suggesting that genres not only are "ever-changing" but that these changes are also discontinuous. Underwood astutely acknowledges that "the things we call genres may be entities of different kinds, with different cycles and degrees

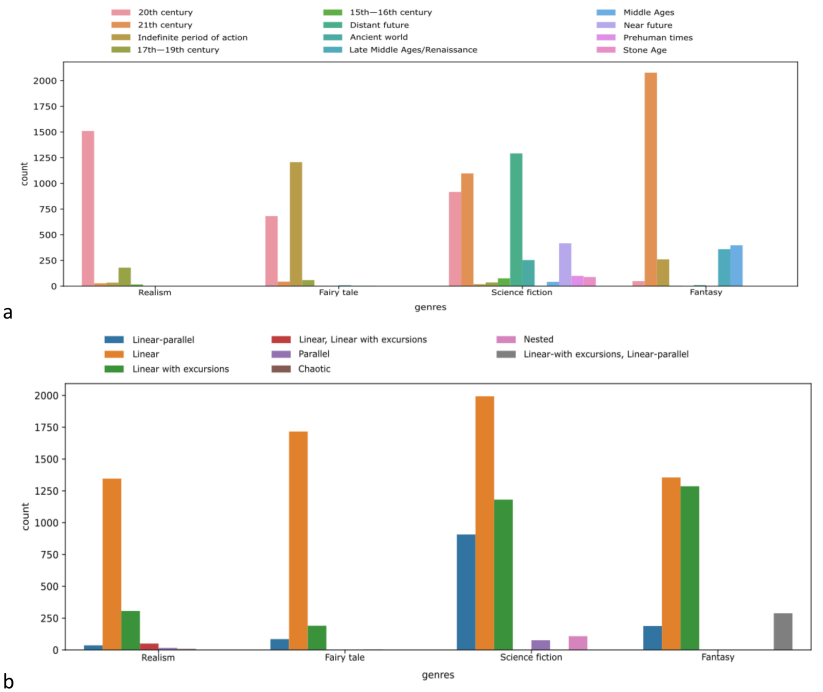

Fig. 4. a) Distribution of plot types by genres; b) Distribution of action times by genres.

of textual coherence" ("Life Cycles" 24). In this regard, we find ourselves in agreement with his observations. That might explain why different methods are better suited for different "entities." For instance, a classifier based on topic modeling would be quite effective for determining texts of science fiction, given its rich and specific lexicon. However, for fairy tales, its performance is barely above random guessing.

Readers' classification of literary works on specialized forums might result in quite diverse works being labeled under, for instance, "fairy tale." In these classifications, folklore-based fairy tales find themselves juxtaposed with more contemporary literary authors' fairy tales, which can exhibit significant differences, such as the inclusion or absence of fantastic elements. Similarly, Roman Jakobson observed that when readers use the term *realism*, they are not solely referring to a specific nineteenth-century literary school but rather something that appears "plausible" (19).

It should be noted that in *FantLab* data, genre categories are not mutually exclusive; for our analysis, however, we selected the most voted label for each

text. For instance, while James Gifford regards sci-fi as a "realistic" genre, in our analysis, we categorized realism and science fiction into distinct genres. Despite this, the quantitative measurement of semantic similarity between groups is weak to moderate. The question of intragenre coherency and its consistency over time requires further exploration, and the results may vary for different "entities." For instance, we observed significant diversification in what is perceived as a fairy tale across the twentieth century, while the young genre of fantasy appears to be relatively coherent. The explored categories may be "entities of different kinds," greatly influenced by the passage of time.

Limitations and Discussion

This exploratory study has several limitations. First, the match of collected data with *Detcorpus* is relatively small and imbalanced. Second, we used only a limited amount of information for training the genre classifier. Most notably, we did not explore the role of subgenres. We also refrained from trying to isolate the author's influence from the genre's influence, considering that numerous authors contribute to the same genre. These concerns suggest being careful when extending our findings beyond the context covered by our data.

Our results show that it is possible to predict the genre of a text with relatively good quality based on its word embeddings. However, we observe that semantic embeddings, although a rich representation of text, may not be enough by themselves. One direction for future work is to explore classifiers of other user-assigned features closely linked to genre as revealed by effect sizes and to see how they perform together in ensembles for genre classification. Another direction, provided by our insightful reviewer, is to incorporate the dimension of authorship into the analyses.

Another direction for future research is to look at the process of readers assigning genre as time series data. Genre boundaries, it seems, are path-dependent: genre assignments in the past shape genre assignments in the present. This observation stresses the historicity of genre categories and the potential effects of large-scale cultural and social processes on boundaries between genres.

Data and Code Availability

Code, FantLab data, and auxiliary reports are available at the following repository: https://osf.io/udvf3/. Due to copyright restrictions, we are unable to share the Detcorpus data. At request, we can share document embeddings to ensure reproducibility.

Yauheniya Lekarevich is a PhD candidate at International Graduate Centre for the Study of Culture at Justus Liebig University Giessen. She is a member of the editorial board of Children's Readings journal. Her dissertation project focuses on the computational analysis of the interaction of age and agency in children's literature written in the Russian language.

Sergei Pashakhin is a PhD researcher with the Chair of Political Science, with a special focus on digital transformations at the University of Bamberg. Their teaching and research focuses on digital platforms, autocratization, computational methods for text analysis, and machine learning.

Notes

[1] For comprehensive details regarding the composition of *Detcorpus*, please refer to the work by Maslinsky et al.

[2] Data were gathered in 2021; refer to "Data and Code Availability" for the list of analyzed texts.

[3] Interactive versions of figures 1–3 can be accessed in the article's repository (refer to "Data and Code Availability"). These versions enable exploring all titles of visualized texts. However, in the print version, full annotations would not be legible.

Works Cited

Adukia, Anjali, et al. "What We Teach about Race and Gender: Representation in Images and Text of Children's Books." *Quarterly Journal of Economics*, vol. 138, no. 4, Nov. 2023, pp. 2225–85.

Antoniak, Maria, et al. "Tags, Borders, and Catalogs: Social Re-working of Genre on LibraryThing." *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, article 29, 2021, pp. 1–29.

Ardanuy, Mariona Coll, and Caroline Sporleder. "Structure-based Clustering of Novels." *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, Association for Computational Linguistics, 2014, pp. 31–39.

Blei, David M., et al. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, vol. 3, Jan. 2003, pp. 993–1022.

Calvo Tello, José. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld UP, 2021.

Du, Keli, et al. "Evaluation of Measures of Distinctiveness. Classification of Literary Texts on the Basis of Distinctive Words." *JCLS*, vol. 1, no. 1, 2022. https://jcls.io/article/id/102/.

Duff, David, editor. *Modern Genre Theory*. Routledge, 2014.

Dunst, Alexander, and Rita Hartel. "The Quantitative Analysis of Comics: Towards a Visual Stylometry of Graphic Narrative." *Empirical Comics Research*, edited by Alexander Dunst, Jochen Laubrock, Janina Wildfeuer, Routledge, 2018, pp. 43–61.

Fernandez, Alberto, et al. "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary." *Journal of Artificial Intelligence Research*, vol. 61, Apr. 2018, pp. 863–905.

Freidenberg, Olga [Фрейденберг, Ольга]. Поэтика сюжета *и* жанра [*Poetika suzheta i zhanra; Poetics of Plot and Genre*]. Labirint, 1997.

Frow, John. *Genre*. Routledge, 2014.

Jakobson, Roman [Якобсон, Роман]. "О художественном вымысле" ["O hudozhestvennom vymysle"]. *Readings in Russian Poetics*, edited by Ladislav Matejka, Department of Slavic Languages and Literatures, 1971, pp. 19–28.

Gifford, James. *A Modernist Fantasy: Modernism, Anarchism, and the Radical Fantastic*. E-book ed., ELS Editions, 2018.

Gittel, Benjamin. "An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500-2020." *Journal of Cultural Analytics*, vol. 6, no. 1, 2021, pp. 217–54.

Gubar, Marah. "On Not Defining Children's Literature." *PMLA*, vol. 126, no. 1, 2011, pp. 209–216.

Hettinger, Lena, et al. "Genre Classification on German Novels." *26th International Workshop on Database and Expert Systems Applications*, IEEE, 2015, pp. 249–53.

Hunt, Peter, ed. *International Companion Encyclopedia of Children's Literature*. Routledge, 2004.

Hunt, Peter, ed. *Understanding Children's Literature*. Routledge, 2006.

Kovtun, Elena. "Science Fiction and Fantasy: Competition and Dialogue in Modern Russia." *Moscow University Bulletin*, series 9, vol. 4, 2015, pp. 53–71.

Le, Quoc, and Tomas Mikolov. "Distributed Representations of Sentences and Documents." *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, no. 2, 2014, pp. 1188–96.

Leiderman, Naum [Лейдерман, Наум]. Теория жанра [*Theoriya zhanra; The Genre Theory*]. Slovesnik, 2010.

Maslinsky, Kirill. "How Exactly Does Literary Content Depend on Genre? A Case Study of Animals in Children's Literature." Computational Humanities Research, 2023, discourse.computational-humanities-research.org/t/how-exactly-does-literary-content-depend-on-genre-a-case-study-of-animals-in-children-s-literature/2048.

Maslinsky, Kirill [Маслинский, Кирилл]. "Сто лет счастья в детской литературе (1920–2020): сталинский канон и его долгосрочные последствия" [Sto let schast'ya v detskoj literature (1920–2020): stalinskij kanon i ego dolgosrochnye

posledstviya; One Hundred Years of Happiness in Children's Literature (1920–2020): The Stalinist Canon and Its Long-term Consequences]. *Steps*, vol. 8, no. 4, 2022, pp. 226–47.

Maslinsky, Kirill [Маслинский, Кирилл], et al. Корпус русской прозы для детей и юношества [*Korpus russkoj prozy dlya detej i yunoshestva; Corpus of Russian Prose for Children and Youth*]. *Open Data Repository for Russian Literature and Folklore*, V2, 2021, https://doi.org/10.31860/openlit-2021.4-C001.

Tynyanov, Yuri [Тынянов, Юрий]. [O literaturnoj evolucii]. *Readings in Russian Poetics*. Edited by Ladislav Matejka, Department of Slavic Languages and Literatures, 1971, pp. 99–114.

Nikolajeva, Maria. "Children's Literature as a Cultural Code: A Semiotic Approach to History." *Aspects and Issues in the History of Children's Literature*, edited by Maria Nikolajeva, 1995, pp. 39–48.

Nodelman, Perry. *The Hidden Adult: Defining Children's Literature*. Johns Hopkins UP, 2008.

Oktyabrskaya, Olga [Октябрьская, Ольга]. Формирование и развитие жанровой системы в русской детской литероатуре *1920–50-х* годов [*Formirovanie i razvitie zhanrovoj sistemy v russkoj detskoj literoature 1920–50-h godov; Formation and Development of the Genre System in Russian Children's Literature of the 1920s–50s*]. 2017. Moscow State U, PhD dissertation.

Ozsarfati, Eran, et al. "Book Genre Classification Based on Titles with Comparative Machine Learning Algorithms." *4th International Conference on Computer and Communication Systems*, IEEE, 2019, pp. 14–20.

Pramesti, Tri. "Considering Young Adult Literature as a Literary Genre." *Parafrase: Jurnal Kajian Kebahasaan & Kesastraan*, vol. 15, no. 1, 2015, pp. 1–10.

Rahul, Ayush, et al. "Genre Classification Using Character Networks." *5th International Conference on Intelligent Computing and Control Systems*, IEEE, 2021, pp. 216–22.

Reynolds, Kimberley. *Children's literature: A Very Short Introduction*. E-book ed., Oxford UP, 2011.

Schöch, Christof. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *DHQ*, vol. 11, no. 2, 2017. https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Sethy, Abhisek, et al. "Book Genre Classification System Using Machine Learning Approach: A Survey." *Proceedings of 5th ICSCSP*, edited V. S. Reddy et al., Springer, 2022, pp. 231–39.

Sobchuk, Oleg, and Artjoms Šeļa. "Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction." *ArXiv*:2305.11251, 2023. https://doi.org/10.48550/arXiv.2305.11251.

———. "(Distant) Reading and (Cultural) Evolution." *NLO*, no. 2, 2018, pp. 88–98.

Tamarchenko, Natan [Тамарченко, Натан], editor. Теория литературных жанров [*Teoriya literaturnykh zhanrov; Theory of Literary Genres*]. 2nd ed., Akademiya, 2012.

Todorov, Tzvetan. "The Origin of Genres." *Modern Genre Theory*, edited by David Duff, Routledge, 2014, pp. 193–209.

Underwood, Ted. "Genre Theory and Historicism." *Cultural Analytics*, vol. 2, no. 2, Oct. 2016. https://culturalanalytics.org/article/11063-genre-theory-and-historicism.

———. "The Life Cycles of Genres." *Cultural Analytics*, vol. 2, no. 2, May 2017, https://culturalanalytics.org/article/11061-the-life-cycles-of-genres.

Walsh, Melanie, and Maria Antoniak. "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." *Journal of Cultural Analytics*, vol. 6, no. 2, 2021, pp. 243–87.

Winner, Thomas. "Structural and Semiotic Genre Theory." *Theories of Literary Genre*, edited by Joseph Strelka, Pennsylvania State UP, 1978.

Yarkho, Boris [Ярхо, Борис]. Методология точного литературоведения [*Metodologiya tochnogo literaturovedeniya; Methodology of Precise Literary Studies*]. Languages of Slavic Cultures, 2006.