

How Exactly does Literary Content Depend on Genre? A Case Study of Animals in Children’s Literature

Kirill Maslinsky^{1,2}

¹INALCO, Paris

²Institute of Russian Literature (Pushkin House), Saint Petersburg

Abstract

The content of literary fiction at least partly depends on literary tradition. The dependence is attested quantitatively in the association of genre with lexical statistical patterns. This short paper is a step to formal modeling of the content-moderating processes associated with literary genres. The idea is to explain prevalence of the particular lemmas in a literary text by the genre-dependent accessibility of the semantic category during the creative process. Data on animals mentioned in various sub-genres in a corpus of Russian children’s literature is used as an empirical case. Vocabulary growth models are applied to infer genre-related differences in overall diversity of animal vocabularies. A constrained topic model is employed to infer preferences for particular animal lemmas displayed by various genres. Results demonstrate the models’ potential to infer genre-related content preferences in the context of high variance and data imbalance.

Keywords

computational thematics, genre, vocabulary growth model, children’s literature

1. Introduction


Computational methods made content of literary fiction a practical target for systematic exploration. All kinds of phenomena are being counted in corpora of fictional texts, including natural and material objects, body parts, emotions etc. with inferences for either literary or cultural history [5, 18, 12]. This body of work could benefit from a more explicit recognition of the dual source of the literary content: literary tradition (internal) and social reality mediated by the author’s experience (external) [19]. It is crucial for inferences on cultural dynamics with literary data to acknowledge and to measure the influence of literary tradition on fictional content.


There is evidence that some aspects of literary tradition captured by a vague notion of genre leave a discernible signal in the distribution of content words. For instance, predictive models using only frequent lexical features are able to discriminate between literary genres with decent accuracy [17, 11, 14, 3, 15]. Such models could be reverse-engineered to look for the most informative lexical features. Yet interpreting these features without understanding the

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

✉ kirill.maslinskii@inalco.fr (K. Maslinsky)

ORCID 0000-0002-9674-2046 (K. Maslinsky)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

mechanics of how genre-related constraints on content translate into lexical distribution phenomena is a risky business. Briefly, quantitative studies of literary content are in need of a more formalized theory.

This short paper is a step to formal modeling of the content-moderating processes associated with literary genres. To reduce complexity, I focus on a narrow subject — animals in children’s literature. The presence of animals in books for children is evidently supported by literary tradition, not only as characters, but more generally as pedagogical material [4, 13]. It is also reasonable to expect variation by sub-genre in the prominence and selection of animals, compare e.g. fairy tales and teen detective stories. Hence I start from the premise that in this case the influence of literary tradition is considerable, is associated with genre, and thus could be measured. The goal is to devise the most simple yet justifiable generative models that represent the content of literary work as a result of choices made during the creative process conditional on genre.

The models suggested in this paper are employed to make two types of measurement in a corpus of Russian literature for children and young adults. First, a quantitative estimation of the effect of sub-genre on the number of various animals mentioned in a text (animal diversity). I suggest to base this estimate on a vocabulary growth model to effectively control for the highly variable text length. Second, an estimation of the preferences to mention certain animal species in each sub-genre. This task is attained with the help of a specialized topic model.

2. Models

To quantify the relative weight of literary tradition vis-a-vis external factors in the prevalence of a concept one needs to put heterogeneous internal and external influences on a unified numeric scale. To this end I suggest to employ a notion of cognitive *accessibility* of a concept to the author in the process of writing. Accessibility can be operationalized as a probability that a concept will be mentioned at least once in a certain work given all the predictive factors. Accessibility of a concept during the creative process is not directly observable, at least not for past literature. But it can be measured *a posteriori* at the population level by observing lexical frequency. Then all the literary and social factors can be seen as distal causes that exert their influence on literary content through increased or decreased accessibility of some concepts. Accessibility provides a convenient conceptual basis for the following models since it offers both a generative interpretation and a measurement scale for the data on lexical prevalence.

2.1. Vocabulary growth model

The first objective is to estimate relative accessibility of animals in general in various sub-genres. The task is complicated by the fact that the distribution of text lengths varies widely between genres. From a modeling perspective the task is to predict the length of a list of different animals mentioned in the text. Since such a list is technically just a part of the text’s vocabulary, a general vocabulary growth model can be applied to it. The most basic model that relates vocabulary size to the text length is known as Heaps’ law [7]. It reflects the fact that vocabulary size in a text in natural language is unbounded, but the growth rate diminishes with

text length. So far as the model is targeted only at a share of the total vocabulary, I slightly modify the interpretation of coefficients in the Heaps' formula

$$V = kT^b \quad (1)$$

where V is the number of animals mentioned (vocabulary size), T is text length in tokens, b (typically $0.4 \leq b \leq 0.6$) and k – coefficients that control the growth rate. This model allows to account for the length of a text in a principled way.

In the experiments below I explore two ways to incorporate the effect of genre into this model. The most obvious move is to allow either the coefficient k or the exponent b to vary by genre. Higher values of the coefficients would indicate higher accessibility of a lexical category in the genre. Evidently, genre (as a proxy to literary tradition) is only partly responsible for the lexical content of the work, and much genre-internal variation remains to be explained by other factors. The external factors that span all the aspects of socialization and linguistic experience of the author can be accounted for by letting either k or b to vary by author. However this solution entails the assumption that authors employ animal vocabulary to a similar degree in all their texts.

The alternative view is to (simplistically) assume that the observed list of animal mentions comes from either of two processes: (a) a low-intensity background process when the number of animals mentioned grows only slowly with the text length; (b) a high-intensity foreground process, leading to a higher number of animal mentions for the text of a similar length. In other words, animals may or may not be a relevant *topic* for a given text. Then each genre could be represented as a mixture of texts each one coming from one of the two processes. The genres would differ in the latter model by the estimated share of texts produced by background and foreground processes. Formally,

$$V = p_g k_1 T^b + (1 - p_g) k_2 T^b \quad (2)$$

where p_g is a genre-specific share of background process, k_1 and k_2 stand for the intensity of a background and foreground processes, respectively.

For details on formal definition of the statistical models used for the experiments, priors, and model selection see appendix A.

2.2. Genre-topic model

The second objective is to estimate the relative accessibility of certain animal species in various sub-genres. For this case, the list of different animals mentioned in a text is treated as a document. The theoretical assumption is that items in this list are drawn from two sources: influence of the literary tradition (genre) on the one side, and the external author's experience on the other. A topic model is the most common formal generative model to describe the composition of a word list drawn from various sources. An advantage of a topic model is that it implicitly accounts for the text length.

The two sources assumption can be translated into a highly constrained topic model where each document is composed of just two topics, one topic specific for the genre of the text, and another topic to model external influences. As a result, each genre has its own "topic". Probabilities of words in these topics reflect preference (higher accessibility) for an animal associated

with a particular genre. While genre-specific topics are meant to capture the influence of literary tradition, for simplicity and to make estimation possible I reduce all external factors to a single common topic.

The generative story for this model runs as follows.

1. For each document d :
Draw a proportion of genre-specific topic p_d for this document.
2. For each word w in a document:
 - a) With probability p_d draw a word from a genre-specific topic $w|z_g$.
 - b) With probability $1 - p_d$ draw a word from a general topic $w|z_c$.

The model has two hyperparameters: a prior for genre-topic proportion in a document $p_d \sim \text{Beta}(3, 3)$ and a Dirichlet prior for distribution of probability of words in each topic $p(w|z) \sim \text{Dirichlet}(\beta_1, \dots, \beta_V), \beta_n = 0.8$.

Such a model can be seen as a highly constrained variant of a well-known LDA topic model [1]. Unlike LDA, the topical composition of a document is not a parameter to be estimated by the model, but is always a mix of two topics pre-defined by the genre of a text. The only document-level parameter the model is left to estimate is a proportion of genre topic. The probabilities of words in a topic are estimated in the same way as in LDA.

3. Data and measurements

The data for the analysis come from the Detcorpus, a corpus of Russian prose for children and young adults written between 1900 and 2020 [9]. All the texts in the corpus are provided with a list of genre tags as a part of their metadata. The genres considered in the present analysis do not form a neat typology. The major genres that span the whole corpus include fairy tale, science fiction, and realism, the last one generic, standing for all prose without specific genre attributes. The other group is formed by formulaic genres that appeared on the market since 1990s: detective stories, fantasy, horror, and romance books for teens. Animal stories, a well-recognizable sub-genre of prose for children is included as a separate category due to a specific focus on animals. For each work, genre tags were reduced to one single label from the above list. Genre labels are regarded as a proxy to those aspects of literary tradition that supposedly have a sufficiently strong and stable effect to be detected in the distribution of animal mentions. In total, the data comprises 2994 works ranging from 100 to 300000 words in length. See details on the data composition, genre and author distribution in the appendix C.

To identify the occurrences of animals in texts I constructed a dictionary using all Russian names and aliases for animal taxa in Wikidata. In contrast to previous work that aimed to measure biodiversity in literature [6, 10], taxa names are not reduced to nouns, and when a taxon name is a multi-word expression it is matched as a sequence of lemmas. Dictionary-based methods are notoriously plagued by false positive matches due to homonymy. The problem is quite severe with animal names, as metaphor is heavily used as a semantic device in this lexical category. To achieve a satisfactory precision, I manually compiled an extensive stoplist (405 items) of the lemmas that are less likely to refer to animals *in this particular corpus*. As a result, of 20811 lemmas in the dictionary 1906 were matched in the corpus. The accuracy of

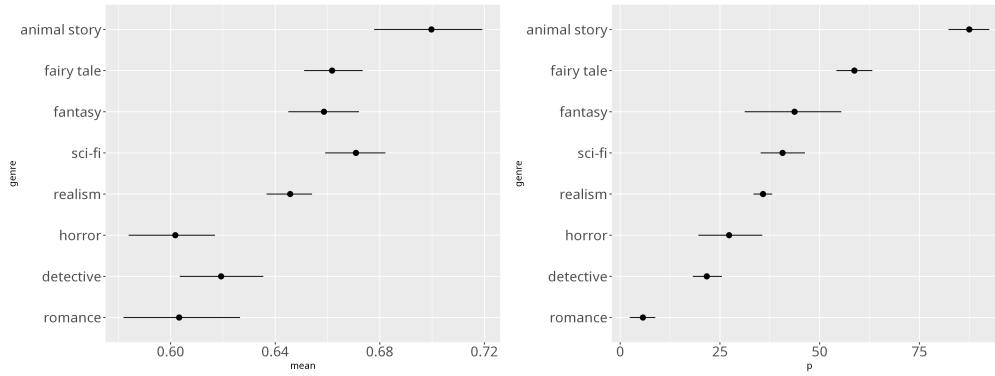


Figure 1: Posterior estimates (mean and 89% compatibility interval) for parameters of vocabulary growth models. Left: growth rate exponent b_g for model (1). Right: percentage p_g of animal-rich texts for model (2)

the method was evaluated on a sample of 50 random 500-word excerpts (precision 0.97, recall 0.81, F1 0.88). Evaluation indicated that Wikidata systematically underrepresents female forms of animal names, names for cubs, and various derivative forms, especially diminutives.

For modeling, each work is reduced to a list of all distinct animal lemmas mentioned in the text (each lemma is present only once). Since the focus of the analysis is on lexical phenomena, I chose to count lemmas, not species. It should be recognized that a relationship of animal nominations to biological taxa can be highly ambiguous, and identification of the specific taxon meant in a text presents a separate problem. For a genre-topic model all the works by the same author in the certain genre are joined into a single document to avoid bias induced by better represented authors in the corpus. To simplify topic inference, only the lemmas that occurred in 5 works or more were retained. All statistical inference was Bayesian and performed with the help of STAN Hamiltonian Monte Carlo sampler. See [8] for the data and code used for the analysis.

4. Results

In the first experiment vocabulary growth models defined above were employed to quantify genre differences in the expected number of animals mentioned (diversity of animal vocabulary). The results are displayed in the fig. 1. Left panel shows the animal vocabulary growth rate parameter b as estimated by the model that assumes that k in the Heaps' formula varies by author and b varies by genre. Right panel displays the percentage p of animal-rich texts for each genre as estimated by a mixture model. Growth rate parameters k and b in the mixture model are fixed for rich and small animal vocabulary clusters and do not depend on author or genre.

Both models infer similar genre differences. Animal stories and, to a lesser degree, genres with fantastic element (fairy tales, fantasy, sci-fi) have larger animal vocabularies on average (or, alternatively, larger share of texts with rich animal vocabularies) in comparison to realism as a reference point. A slightly more surprising conclusion is that formulaic genres (detective,

horror) use narrower animal vocabulary with the lowest result attained by teen romance novels.

The vocabulary growth rate model indicates that more variance in the animal vocabulary size is associated with authors than with genres. The model predicts that in a typical novella (50,000 words) an author with an average interest in animals will mention 10 more animal lemmas, on average, in an animal story, 3 more in a fairy tale, and 6 less in romance, all in comparison with realism. Simultaneously, the predicted difference between the author with the highest interest in animals (Nikolai Sladkov) with one of the lowest (Anatoly Aleksin) for a realist novella of the same length would be 197 animal lemmas, on average.

Since the mixture model estimates a probability that each text belongs to either rich or small animal vocabulary group it could be seen as a model-based clustering of texts. This allows for finer comparison of otherwise similar works that differ by the density of animal mentions. Many authors consistently appear in one of the clusters, for instance, Vitaliy Bianki, a canonical author of animal stories, is invariably a high-scorer. But even the texts of the same genre, size and by the same author may fall in different groups. Short stories from the same book by Andrei Platonov classified as fairy tales provide a vivid example. One is “Why did the geese become motley”, 626 words, 1 animal species (geese), low-animal cluster. The second is “A grateful hare”, 643 words, 11 species, high-animal cluster. In the second story, a hare helps the protagonist by calling other animals to bring foods, which effectively generates an enumeration of species.

For the second experiment a genre-topic model was trained. The parameters estimated by the model include the probabilities for each lemma in each genre (genre-specific topics) and a background probability of each lemma. The probability distribution of lemmas in the background topic is very close to the overall frequency distribution of lemmas in the data (Jensen-Shannon divergence 0.02). High probability of a lemma in a genre topic means that this animal is likely to be mentioned by larger number of authors writing in this genre (is more accessible given the genre). The summary of the genre topics is presented in fig. 2. For generality, instead of presenting lemmas I group animals in larger categories and provide a number of lemmas in each category for a top-20 lemmas in a topic. Top lists for each topic are built using a balanced FREX metric which combines probability of a lemma in a topic with exclusivity of a lemma to this topic in contrast to other topics.

There are a few notable tendencies made apparent thanks to the genre-topic model. Animal stories are distinctive for its focus on forest animals and the most diverse set of bird species, primarily wildfowl. This may be contextualized with a note that the most prolific authors in this genre (e. g. Bianki, Prishvin) were passionate hunters. Birds also have a prominent place in other genres, including fairy tales, but a set of species is quite different (various owls, crow, tit). In contrast to animal stories, realism is defined by its focus on farm animals (horse being the most “realistic”) and quite numerous species of edible fish. Perhaps not surprisingly, interest in snakes is characteristic of teen horror fiction. Science fiction is much more focused on sea animals along with extinct species (often dinosaurs). Pets (primarily cats and dogs) take a very prominent place in detectives. All animals not native for northern temperate zone are grouped under a label ‘exotic’. For instance, lion (‘king of animals’) and tiger are typical of fairy tales.



Figure 2: Distribution of top-20 lemmas for each genre topic by animal class

5. Discussion

The models introduced in this short paper aim to ground computational analysis of literary content (or computational thematics, as suggested by Sobchuk and Šeĭa in [15]) in the categories relevant for literary production. While the models are simplistic they operate on a level of a whole literary work which allows to relate them to the aspects of the creative process. The animal diversity model was able to detect genre-associated differences in the accessibility of animal vocabularies while controlling for the author and text length. This result corroborates the supposed effect of literary tradition for this particular data. The mixture model that distinguishes between rich and scarce animal vocabularies also proved to be a useful tool to locate diversity-generating tropes, as shown in the Platonov example in section 4.

I see an important advantage of the vocabulary growth models employed here in their ability to estimate parameters even for the very short texts on a par with longer ones. This feature is specifically relevant for any diversity measurements made on a lexical basis, say, biodiversity as represented in literature. The reason is that text length is the strongest predictor of the vocabulary size regardless of other factors [16]. For this reason previous work on literary biodiversity [6, 10] had to recourse to the minimum text length threshold which is equivalent to implicitly stratifying by text length without a proper theoretical justification.

In comparison to linear prevalence models (Poisson regression) vocabulary growth models

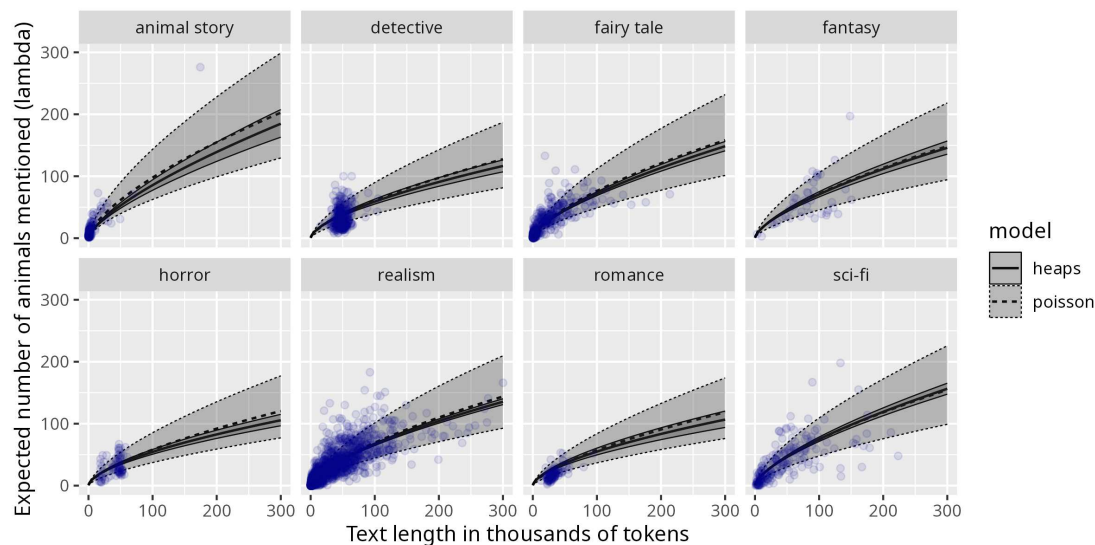


Figure 3: Expected size of animal vocabulary in various genres, for an author with an average interest in animals. The posterior means and 89% compatibility intervals are displayed for a Heaps’ vocabulary growth model and Poisson generalized linear model. Overlaid data points display size of animal vocabulary in each work in Detcorpus

may offer more fine-grained estimates. For instance, given a point estimate of the author’s propensity to mention animals, both the Heaps’-based model suggested in this article and a similarly structured Poisson model (see appendix B for details) produce rather similar posterior inferences. But in case of a Poisson model, the parameter estimate uncertainty grows quickly with the text length, unlike the Heaps’ model (fig. 3).

An important limitation of the vocabulary growth models based on the Heaps’ formula in comparison to linear models is that various predictors cannot be so easily incorporated into the model. While coefficients k and b provide two points to stratify by predictor variables, their effects on the outcome are not symmetric. Moreover, adding more than two predictors (for instance, as factors of k) runs into an identification problem in the context of Bayesian inference. Reparameterizing the model or switching to another basic vocabulary growth model may be required to tackle this problem.

The genre-topic model that captures preferences for specific animal lemmas conceptually describes the animal profile of the genre as a deviation from the corpus-wide distribution of animal frequencies. This is structurally analogous to a popular idea in stylometry that is behind the Burrows’ Delta [2] and was shown to work for the genre classification as well [14, 15]. The advantage of the Bayesian genre-topic model in comparison to simpler measures of lexical distinctiveness is that it provides estimates for the uncertainty of the parameters. Highly uncertain estimates for the proportion of the genre-related topic in a document indicates that with this particular model and data it is not possible to tell to what extent the animal vocabulary of a given author is defined by the literary tradition or by the external factors. Nevertheless, the

model was able to detect relatively minor genre-related modulations in frequency of certain animal species in the presence of a strong signal of a general linguistic or literary background.

References

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [2] J. Burrows. “Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship”. In: *Literary and linguistic computing* 17.3 (2002), pp. 267–287.
- [3] J. Calvo Tello. *The Novel in the Spanish Silver Age*. Wetzlar: Bielefeld University Press, 2021.
- [4] T. Cosslett. *Talking animals in British children’s fiction, 1786–1914*. New York: Routledge, 2017.
- [5] R. Heuser and L. Le-Khac. *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*. Pamphlets of the Stanford Literary Lab 4. 2012. URL: <http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf>.
- [6] L. Langer, M. Burghardt, R. Borgards, K. Böhning-Gaese, R. Seppelt, and C. Wirth. “The Rise and Fall of Biodiversity in Literature: A Comprehensive Quantification of Historical Changes in the Use of Vernacular Labels for Biological Taxa in Western Creative Literature”. In: *People and Nature* 3.5 (2021), pp. 1093–1109.
- [7] D. C. van Leijenhorst and T. P. Van der Weide. “A Formal Derivation of Heaps’ Law”. In: *Information Sciences* 170.2-4 (2005), pp. 263–272. DOI: 10.1016/j.ins.2004.03.006.
- [8] K. Maslinsky. *Replication Data for: How Exactly does Literary Content Depend on Genre? A Case Study of Animals in Children’s Literature*. Repository of Open Data on Russian Literature and Folklore. Version V1. 2023. DOI: 10.31860/openlit-2023.10-R005.
- [9] K. Maslinsky, Y. Lekarevich, and L. Aleinik. *Corpus of Russian Prose for Children and Young Adults*. Repository of Open Data on Russian Literature and Folklore. Version V2. 2021. DOI: 10.31860/openlit-2021.4-C001.
- [10] A. Piper. “Biodiversity is not Declining in Fiction”. In: *Journal of Cultural Analytics* 7.3 (2022). DOI: 10.22148/001c.38739.
- [11] A. Piper. “Fictionality”. In: *Journal of Cultural Analytics* 2.2 (2016). DOI: 10.22148/16.011.
- [12] A. Piper and S. Bagga. “A Quantitative Study of Fictional Things”. In: *Proceedings of the Computational Humanities Research Conference*. Antwerp, Belgium, 2022, pp. 268–279. URL: <https://ceur-ws.org/Vol-3290/long%5C%5Fpaper1576.pdf>.
- [13] H. Ritvo. “Learning from Animals: Natural History for Children in the Eighteenth and Nineteenth Centuries”. In: *Children’s literature* 13.1 (1985), pp. 72–93.
- [14] A. Sharmaa, Y. Hu, P. Wu, W. Shang, S. Singhal, and T. Underwood. “The rise and fall of genre differentiation in English-language fiction”. In: *DH2020 (ADHO) Proceedings*. Amsterdam, 2020, pp. 97–114.

- [15] O. Sobchuk and A. Šeja. “Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction”. In: *arXiv preprint arXiv:2305.11251* (2023).
- [16] F. J. Tweedie and R. H. Baayen. “How Variable May a Constant be? Measures of Lexical Richness in Perspective”. In: *Computers and the Humanities* 32 (1998), pp. 323–352.
- [17] T. Underwood. “The Life Spans of Genres”. In: *Distant horizons: digital evidence and literary change*. Chicago: University of Chicago Press, 2019. Chap. 2, pp. 34–67.
- [18] T. Underwood, D. Bamman, and S. Lee. “The Transformation of Gender in English-language Fiction”. In: *Journal of Cultural Analytics* 3.2 (2018). DOI: 10.22148/16.019.
- [19] B. Yarkho. “Metodologiya Tochnogo Literaturovedeniya. Izbrannye Trudy po Teorii Literatury [A Methodology for a Precise Science of Literature. Selected Works on Literary Theory]”. In: Moscow: Languages of Slavic Cultures, 2006, pp. 247–251.

A. The definition of vocabulary growth models

The central idea is to define the expected size of the animal vocabulary in a given text with the help of the Heaps’ formula $V = kT^b$. To adapt to the fact that vocabulary size is a natural number, the expected value predicted by the Heaps’ model can be treated as a parameter (expected value) for a Poisson distribution. To test the hypothesis that there is an effect of literary tradition on accessibility of the animal vocabulary associated with sub-genre of children’s literature, one needs to stratify animal vocabulary growth rates by genre. Alternatively, inter-genre variance in animal vocabularies may be explained away by external factors (individual author characteristics). Heaps’ formula offers two options to account for genre/author variation: either coefficient k or b could vary by genre or by author.

To select the final model I tested all logically possible combinations of k and b coefficients associated with either genre or author. The best performing model was selected by evaluating the model’s predictive ability for the animal vocabulary data in children’s literature with the help of the WAIC criterion. The model comparison summary is presented in table 1. For animal vocabulary data, coefficient k turns out to be more effective in capturing data variance in comparison to b . It works this way both for author-based variance and genre-based variance. Including author as a factor in a formula always results in a much better fit. Whenever author is present, adding genre results in a relatively small (but non-null) model improvement. This improvement is more pronounced if genre is taken into account as a more effective k coefficient.

The formal definition of the selected model follows below in 3, the rest of the models had

Table 1

Result of the comparison of the vocabulary growth model variants (WAIC)

model description	k	b	WAIC	sd	dWAIC	pWAIC
k varies by author, b by genre	author	genre	22392.5	264.58	0.0	1151.4
k varies by author, no genre effect	author	fixed	22432.0	268.99	39.5	1137.7
k varies by genre, b by author	genre	author	23097.9	290.37	705.4	1070.7
b varies by author, no genre effect	fixed	author	23441.0	290.79	1048.5	1074.3
mixture model, no author effect	low/high	fixed	25935.7	379.70	3543.2	29.2
k varies by genre, no author effect	genre	fixed	33489.0	698.90	11096.5	78.3
b varies by genre, no author effect	fixed	genre	33960.7	695.30	11568.2	84.5

similar structure and priors. All models employed partial pooling on author/genre coefficients.

$$\begin{aligned}
V_i &\sim \text{Poisson}(\lambda_i) \\
\lambda_i &= k_a T_i^{b_g} \\
k_a &\sim \text{Log-Normal}(\bar{k}, \sigma_a) \\
\bar{k} &\sim \text{Normal}(1, 0.7) \\
\sigma_a &\sim \text{Exponential}(1) \\
b_g &\sim \text{Normal}(\bar{b}, \sigma_g) \\
\bar{b} &\sim \text{Normal}(0.5, 0.1) \\
\sigma_g &\sim \text{Exponential}(10)
\end{aligned} \tag{3}$$

where V is animal vocabulary size, λ_i stands for the expected Poisson rate for a text i , and T_i is the length of a text in thousands of tokens. External factors that influence accessibility of animal category are captured by k_a that reflects interest in animals for each individual author. Internal (literary) factors are captured by the exponent b_g that varies by genre. The distributions of both author and genre coefficients are defined by higher-order priors \bar{k} , σ_a and \bar{b} , σ_g .

The mixture model represents vocabulary size as a result of either a low-intensity background process, or a high-intensity foreground process, mixed in a genre-specific proportion p_g . The expected vocabulary size value for both processes is defined by the same Heaps' formula. The formal definition of the model is given in 4.

$$\begin{aligned}
V_i &\sim p_g \text{Poisson}(\lambda_1) + (1 - p_g) \text{Poisson}(\lambda_2) \\
\lambda_1 &= k_1 T^b \\
\lambda_2 &= k_2 T^b \\
k &\sim \text{Log-Normal}(1, 0.7) \\
b &\sim \text{Beta}(5, 5) \\
\text{logit}(p_g) &= \alpha_g \\
\alpha_g &\sim \text{Normal}(0, 1)
\end{aligned} \tag{4}$$

where λ_1 and λ_2 stand for the expected value for animal vocabulary for the background and

the foreground processes, respectively. Similarly, k_1 and k_2 denote accessibility coefficients for both processes.

B. Poisson prevalence model

To provide a comparison with the suggested vocabulary growth model, a more traditional Poisson generalized linear model for lexical prevalence was defined and applied to the same data. The model is designed to maximally fit the structure of the Heaps-based vocabulary growth model. Partial pooling of the author and genre coefficients is employed as well. To optimize inference, the non-centered model parameterization was used. Logarithm of text length is taken into account as an exposure parameter. The formal model definition is as follows:

$$\begin{aligned}
 V_i &\sim \text{Poisson}(\lambda_i) \\
 \log(\lambda_i) &= \bar{\alpha} + \beta_a \sigma_a + \beta_g \sigma_g + \beta_T \log T_i \\
 \beta_a &\sim \text{Normal}(0, 1) \\
 \beta_g &\sim \text{Normal}(0, 1) \\
 \beta_T &\sim \text{Normal}(0, 0.25) \\
 \bar{\alpha} &\sim \text{Normal}(0, 0.25) \\
 \sigma_a &\sim \text{Exponential}(1) \\
 \sigma_g &\sim \text{Exponential}(1)
 \end{aligned} \tag{5}$$

C. Corpus details

Texts in the Detcorpus data come with a list of genre tags assigned based on the bibliographic and contextual data. For the present analysis lists of genre tags were run through a simplification procedure to arrive at a single level for each text. In case there are several genre tags for a text, only one of them is retained. Some secondary sub-genre tags are omitted as a result, for instance, “school novel”. If a list of genres contain a fairy tale tag, the text was always regarded as a fairy tale. If there are several genre tags, the first one in a list is regarded as primary and retained. Several genres with a very sparse representation in the corpus were omitted from the analysis (adventure, biography).

The data contains texts written by 917 authors, with the majority of them (89%) represented in a single genre only. See table 2 for details on author and genre distribution.

As a result, one can see that the composition of the dataset in terms of genres is not in any way a balanced sample. The genre preferences changed with time, and the corpus sample is also somewhat imbalanced diachronically, with some decades represented better than others. The distribution of genres by decade are shown on fig. 4. Some genres are represented better than others, with realism (a “default” genre assigned to texts without a specific genre affiliation) spanning 53% of works included in the corpus.

Table 2

Left: Text distribution by genre. Right: Distribution of authors by number of genre they have written in

genre	n	p
realism	1588	53.0
skazka	450	15.0
detective	349	11.7
scifi	205	6.8
animalistic	164	5.5
love	113	3.8
horror	84	2.8
fantasy	41	1.4

ngenres	n	p
1	813	88.7
2	85	9.3
3	15	1.6
4	2	0.2
5	1	0.1
6	1	0.1

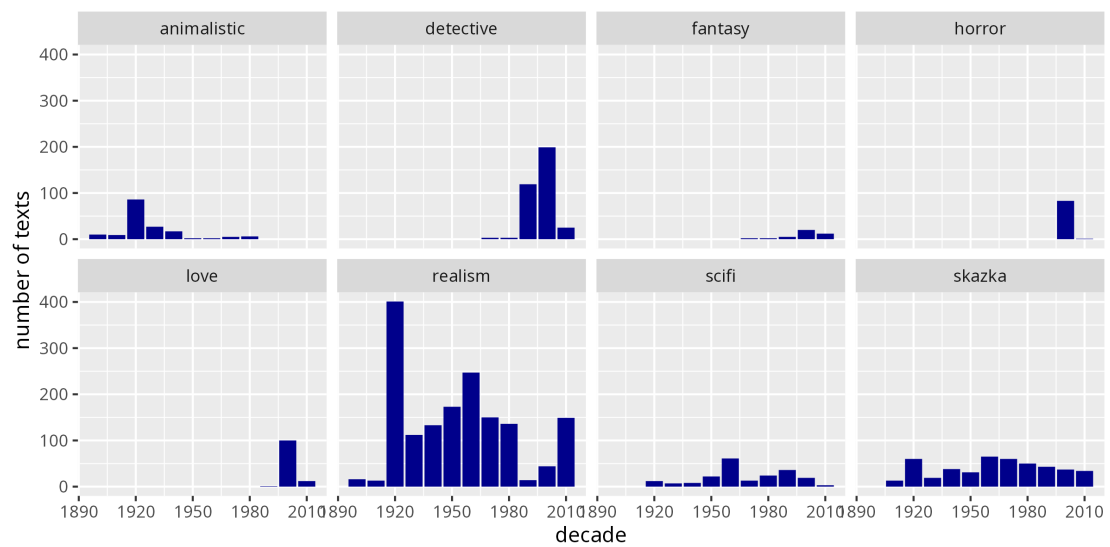


Figure 4: Distribution of genres by decade