http://doi.org/10.15446/lthc.v25n2.108775

Unread, Yet Preserved: A Case Study on Survival of the 19th-Century Printed Poetry

Antonina Martynenko
University of Tartu, Estonia
antonina.martynenko@ut.ee

Distant reading promises access to "the great unread", which should allow scholars to rethink the history of literature. However, the rise in volume of data does not guarantee the understanding of a *corpus* and its relation to the literary population. This article discusses how a "complete" *corpus* of the 19th-century poetry books in Russian might be collected with account for historical data and potential survivorship bias. Even if bibliographical sources cannot provide a complete list of books printed in a given period, the amount of "incompleteness" can be directly estimated with the unseen species models. The estimation of survival ratios for printed poetry shows differences in the loss rate across different types of sources: with conventional editions, like books and anthologies, are well-preserved, while booklets and pamphlets are the largest expected source of loss. These findings allow us to estimate what an "exhaustive" *corpus* can look like and define the features of "the unread" and "unseen" inside it.

Keywords: unseen species problem; computational humanities; poetic *corpus*; 19th century literature; bibliography; book history.

Cómo citar este artículo (MLA): Martynenko, Antonina. "Unread, yet preserved: A case study on survival of the 19th-century printed poetry". *Literatura: teoría, historia, crítica*, vol. 25, núm. 2, 2023, págs. 192-214.

Artículo original. Recibido: 15/01/23; aceptado: 12/04/2023. Publicado en línea: 01/07/2023.

(cc) BY-NC-ND

No leído, pero conservado: un estudio de caso sobre la supervivencia de la poesía impresa del siglo XIX

La lectura distante promete el acceso a "el gran sin leer", lo que debería permitir a los estudiosos repensar la historia de la literatura. Sin embargo, el aumento del volumen de datos no garantiza la comprensión de un *corpus* y su relación con el *corpus* literario. Este artículo analiza cómo se podría recopilar un *corpus* "completo" de los libros de poesía del siglo XIX en ruso teniendo en cuenta los datos históricos y el posible sesgo de supervivencia. Incluso si las fuentes bibliográficas no pueden proporcionar una lista completa de los libros impresos en un período determinado, la cantidad de "incompletitud" se puede estimar directamente con los modelos de especies no-vistas. La estimación de los índices de supervivencia de la poesía impresa muestra diferencias en la tasa de pérdida entre los diferentes tipos de fuentes: las ediciones convencionales, como los libros y las antologías, están bien conservadas, mientras que los cuadernillos y panfletos son la principal fuente esperada de pérdida. Estos hallazgos nos permiten estimar cómo puede ser un *corpus* "exhaustivo" y definir las características de "lo no leído" y lo "no visto" dentro de él.

Palabras clave: problemas de las especies no-vistas; humanidades digitales; *corpus* poético; literatura del siglo XIX; bibliografía; historia del libro.

Não lidos, mas preservados: um estudo de caso sobre a sobrevivência da poesia impressa do século XIX

A leitura à distância promete acesso ao "grande não lido", o que poderia permitir que os acadêmicos repensem a história literária. No entanto, o aumento no volume de dados não garante a compreensão de um *corpus* e sua relação com a história literária. Este artigo discute como um *corpus* "completo" de livros de poesia do século XIX em russo poderia ser compilado levando em consideração dados históricos e possíveis tendências de sobrevivência. Mesmo que as fontes bibliográficas não possam fornecer uma lista completa de livros impressos em um determinado período, a quantidade de "incompletude" pode ser estimada diretamente com modelos de espécies não vistos. A estimativa das taxas de sobrevivência da poesia impressa mostra diferenças nas taxas de perda entre diferentes tipos de fontes: edições convencionais, como livros e antologias, são bem preservadas, enquanto livretos e panfletos são a principal fonte de perda esperada. Essas descobertas nos permitem estimar a aparência de um *corpus* "abrangente" e definir as características do "não lido" e do "não visto" dentro dele.

Palavras-chave: problema das espécies não vistas; humanidades digitais; *corpus* poético; literatura do século XIX; bibliografia; história do livro.

Introduction

One of the most frequent promises of *distant reading* is to reveal previously undiscovered trends and processes in the history of literature by accessing large-scale collections of texts, or so-called "the great unread" . While there is a constant discussion on method application for style and genre analysis, less attention is given to the question of biases involved in the *corpus* creation process. On the one hand, the largest collections as Google Books cover too many materials to be provided with detailed historical metadata and still are incomplete in various ways. On the other hand, carefully curated small corpora are often biased towards literary canon or other selection features. But even a "simple" correspondence of the *corpus* size to the *total* book population size usually cannot be clearly estimated for either of the two. So, what do we study when we study large collections? What is their relationship to the past? How much of "the unread" is in fact missing from our corpora?

This case study addresses the questions of selection and survival bias, taking as an example the process of a poetry *corpus* creation of a narrow historical span (1830–1850). Not comparable with the large-scale collections like Google Books or HathiTrust, poetry as data is usually rather small. Poems are in most cases collected based on some exclusive selection criteria: poetic form (Diachronic Spanish Sonnet *Corpus*), canonical, or "representative" authors (*cf.* Métrique en Ligne, Eighteenth-Century Poetry Archive), or scholarly authority (Poetic sub*corpus* of Russian National *Corpus*).

The latter provides a specifically telling example. It includes quite a significant collection (13M words) supplied with elaborated morphological, metrical, and stanzaic annotation. The selection of authors in the *corpus* was in many cases decided by the availability of modern (the 20th century) scholarly editions (Korchagin 235). As a result, the period of "classical" 19th century poetry in the *corpus* is presented not by primary but secondary sources and affected by Soviet literary canon of the middle of the 20th century (almost all sources of 18–19th century texts were published in the 20th century). Furthermore, the bias towards the most studied poets lead to dramatic overrepresentation of authors like Aleksandr Pushkin and Mikhail Lermontov, who have their complete works and even drafts included in the

corpus, while for the most authors the number of poems in the *corpus* rarely exceed 10–15 texts. This *corpus* design though is not accidental but rooted in the earlier tradition of quantitative yet non-computational versification studies, in particular, these on the history of verse forms by Mikhail Gasparov and his selection of scholarly editions of poetry (Korchagin 235-237).

Without a doubt, the Russian National *Corpus* is a large collection, a "big data" scale of poetic texts. However, the question is simple: given the selection criteria and canonicity bias, how really "unread" is it? Would it be possible to build an inclusive *corpus* from the primary sources without selection? And how many of these primary sources are available for us?

This article aims to examine the path that precede the "exhaustive" *corpus* creation, firstly, by discussing the trends visible from historical circulation and current distribution of primary sources and, secondly, by estimating the *corpus* bibliography completeness with the help of methods from population ecology. I claim that preliminary research into the sources of a digitised collection can bring more understanding of available and even lost data as well as give a clearer vision on what is inside "the unknown" part of a *corpus* and which aspects of literary history "the unread" can disclose.

The structure of this paper is as follows. The Data section gives the historical overview of the data used for this case study. In particular, the section Smirdin's library catalogues describes the path of one library collection through two centuries: from its creation in the 19th century towards its current state; the Contemporary library holdings section summarises the data on preservation of the 19th century poetry books in 11 nowadays library collections. The Poetry books survival rate section presents a methodology for estimation of the books' survival rate and applies the method to the data described in the previous section. The last section discusses limitations of the data and methods used for the case study and outlines further prospects of this approach on a large scale.

Data

The object of this study is Russian poetry of the 19th century, namely the books of poetry issued between 1830 and 1850. To give the necessary context, this period is considered to be the end of the "Golden Age" of Russian poetry, since its political and, partly, aesthetical functions gradually

shifted to prose. The 1830s were already named as a period of poetic decline, marked by the symbolic deaths of Aleksandr Pushkin (1799–1837) and Mikhail Lermontov (1814–1841). The period of the late 1830s and the 1840s was deemed not worthy as a reading and study object.[1] Despite general unfavourable attitude to this period, selected poets active in the 1830s and 1840s entered the 20th century literary canon and 102 of them found their way in the national *corpus*.

A hundred poets might seem like a significant number; however, these two decades were also an active period of professionalisation and democratization of Russian literature: many new poets of different social backgrounds had entered the literary field and tried to push literary activity away from leisure to financially successful endeavour. The transformation started in the middle 1820s with the appearance of an unprecedently popular and profitable literary almanack "The Polar Star" . It introduced high wages per line and caused a flood of copycats. Annual collections of selected poems and short stories that were published in the middle of the 1820s and early 1830s gave name to the period known as "The Epoch of Almanacks" (Smirnov-Sokol'skiy 16-21). Individual books of poetry followed this stream of nice-looking compilations of works by multiple authors. The wave of published poetry was noticed by an unknown critic in 1838:

> How to get away from verses? Where to hide from the inspiration that relentlessly manifest itself in little poetry collections printed on white and grey paper with typos and mistakes in grammar? You take these little books with sweet expectations but then leave them in an unbearable ennui.[2] (Russkaya literatura 791)

The sarcasm on the quality of poetry of this period was a very common point, though the "expectations" in this case might be connected with the idea that publishing a book of poetry is an outstanding act, permitted only for the greatest, recognised authors, as it was in the 1810s and 1820s. Decade later, however, the total cost of issuing a book not only lowered, but even promised some financial profits (Rose & Eliot 343). Numerous novice authors —often completely unknown to their contemporaries and academics

1 See, for instance, Orlov 478-498; Bukhshtab 5-60; Vatsuro 362-379 .
2 All the translations from Russian are mine.

alike— started to publish poetry collections and standalone editions of narrative poems. This is the reason this period is an underexplored source for gathering and examining the "unread".

No reliable bibliography listing the books of poetry published between 1830 and 1850 exists. The archival materials that comprise registers of the books going through censorship are incomplete and hardly available; despite state-controlled recording of all printed books started in 1837, no centralised register for the books printed in Russia had been ever published (Reyser 75-7; Levin 147-53). Literary scholars name the 19th century booksellers' and library catalogues as the most valuable sources for that matter, as well as several contemporary library collections, with the biggest one in the National Library of Russia which was the legal deposit library between 1810 and 1917. To my knowledge, the only attempt to create a list of books of poetry of the 1830s and 1840s belonged to Aleksey Balakin, though his bibliography covered only the period between 1834 and 1850 and was not finished and officially published (Balakin).

Thus, the real number of the allegedly increasing population of poetry books was never quantified, nor was it evaluated if (and how much) these books are still being preserved in the library collections. It can be assumed that despite the low reputation of the poetry of this period, the books might still be preserved because of the high status of poetry itself. Using 19th century indexes and digitalised library catalogues, I will try to assess whether these sources can provide enough data for creating an extended bibliography of poetry books (1830–1850) and for examination of the physical items (re)distribution in contemporary collections. The major goal behind compilation and evaluation of the bibliography is to find whether we can reach an understanding of what is "the unread" in the future *corpus* of poetry, how good it is preserved and which part of the already known poetic texts "the unread" forms.

Smirdin's library catalogues

Among many 19th century catalogues printed by booksellers and bibliographers, the exceptional role belongs to the catalogue of Aleksandr Smirdin's library. Besides the literal value of the catalogues as sources for bibliography, these catalogues also reflect different stages of the collection's

diffusion. That case can be an illustration of how poetry books were redistributed and lost between the 19th and 21st centuries.

For literary historians Smirdin is mostly known as a bookseller and publisher who greatly influenced the development of cheap and profitable bookselling in Russia and established flat fees for authors.[3] He is less remembered for maintaining[4] a commercial library and publishing its catalogues,[5] which later were recognised as one of the most complete (Koblents 82-88) existing bibliographic sources on Russian books printed in the 1820s and 1830s (four-part catalogue appeared in 1828; annexes added in 1829 and 1832). It is especially valuable that the catalogues are organised and allow to immediately find books written in verse. In 1842 the library was passed to Mikhail Ol'khin and then to Petr Krasheninnikov in 1847 (Kishkin 151), both of them also published printed catalogues of the expanded Smirdin's collection (in 1846 and 1852, respectively).

By the general rules of Smirdin's library, the catalogues were used by readers mainly for ordering books as well as for learning prices in case of purchasing "no matter how many items" ("Rospis' …" II-III). The same functionality applied to later catalogues by Ol'khin and Krasheninnikov. Thus, even if the catalogues present exceptionally rich information on books published and read in the 1830s and 1840s, they were not designed as complete bibliographical registers. This limits the direct interpretation of the items not passing from one catalogue to another because a book absence could have been caused by number of factors starting from direct loss (all copies of the book were sold from the library) to other reasons, as a book's low popularity, an editor's mistake or even reasoning that a book was already mentioned in the earlier catalogue. Nevertheless, these three library catalogues issued in 1832, 1846, and 1852 presents the different stages of extension of the same library, giving almost a live broadcast on books of poetry read and sold from 1830 to 1850.

In the second half of the 19th century the collection stopped functioning as a library and was resold[6] several times, until it had finally landed in Nikolay

3 See, for example, a book by Nikolay Smirnov-Sokol'skiy *Knizhnaya lavka Smirdina* [Smirdin's Bookshop] (1957).

4 The library was established by Vasiliy Plavil'shchikov in 1815, in 1823 it was inherited by Smirdin and reopened in the centre of Saint-Petersburg in 1832 (Kishkin 151).

5 Smirdin was probably one of the authors of the early catalogues (Koblents 90-91).

6 The exact route of the library from Saint Petersburg to Riga is unknown, while several

Kymmel's bookshop in Riga, in the 1870s. Kymmel, in turn, also published a number of catalogues aiming to sell the books as rare and valuable parts of *Smirdin's* collections: the bookseller used the library's founder name as a part of the catalogues' title: "Antiquarian catalogue n.º 39: Comprising the list of books from Smirdin's library, nowadays belonging to N. Kymmel, a bookseller in Riga" (1889). To some extent Kymmel even used the system established by Smirdin's catalogues, separating poetry and fiction in two neighbouring catalogues (n.º 39 and 40 respectively, both issued in 1889).[7] The primary interest of this study is the catalogue n.º 39, that reflects the shape of the collection in the late 1880s.

It is worth mentioning that not long before Kymmel's catalogues publication, a part of Smirdin's collection was bought by a bibliophile Pavel Shchapov whose collection is currently kept in the Russian State Public Historical Library ("Knizhnaya kollektsiya…"). According to the lists of books that are now preserved in the library, Shchapov focused on the early 19th century – only five items of the poetry books he bought were printed after 1830. Many others remained in Kymmel's bookshop for sale.

It seems that Kymmel was successful in selling Smirdin's collection, since the next and the final owner of Smirdin's library received only a small part of it. In 1932 the collection was bought and transferred to the Slavonic library of the National library of the Czech Republic; compared to the 1850s, only 10 to 20 percent of volumes reached Prague (Savitskiy 96-97). The least preserved were humanities-related books, especially the sections of history, fiction and, finally, poetry, where "out of 82 registered [in some part of Smirdin's catalogue] items were received only 13" (Savitskiy 97). Nevertheless, the library tried to replenish Smirdin's collection using its older catalogues and nowadays this collection can be considered one of the richest and also digitised collections of 19th century books in Slavic languages.

Incarnations of Smirdin's library allow us to follow the collection's path from its initial state in the 1830s to the present restored condition. I collected data on all books of poetry[8] published between 1830 and 1850 and mentioned

booksellers are believed to have bought parts of the library ("Knizhnaya kollektsiya…").

7 Altogether Kymmel compiled six catalogues based on Smirdin collection, these are catalogues numbered as 33 ("Theology and philosophy", 1887), 34 ("Law, politics and political-economic sciences" , 1887), 36 ("History and geography", 1887), 44 ("Military and marine sciences", 1890) as well as poetry and prose fiction lists 39 and 40 mentioned above.

8 All lyric and narrative genres were included. Drama in verse, folksongs and children's

in the four 19th century catalogues (shortened as Smirdin 1832, Ol'khin 1846, Krasheninnikov 1852, Kymmel 1889) and one modern library database (online catalogues of the Slavonic library). Figure 1 presents the number of shared books between each of the catalogues. It must be noted that the figure reflects only the shared number of books between each two catalogues but by no means displays the complete path of each book from one collection to another (the data derived from the catalogues is not enough for this tracing).

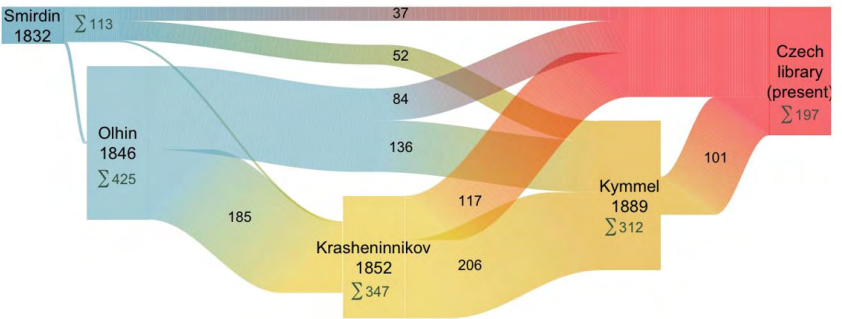

Figure 1. The number of shared books by each of the catalogues presenting Smirdin's library in different time periods (only poetry books issued between 1830 and 1850 included)

It is striking that the original catalogue almost does not share any poetic books with its direct successor, but, instead, has the largest intersection with the 1889 catalogue. The latter can include many of earlier books due to the potential customer's interest in the early-nineteenth-century Russian poetry or, moreover, possibly reflecting a shift in the catalogue's function: from day-to-day to antiquarian bookselling. Proportionally-wise Kymmel' s catalogue includes 46 % of the books (1830–1832) mentioned in catalogue of the 1832 and have a bigger overlap of 59 % with chronologically closest list of 1852. However, the proportion of books it shares with 1846 catalogue is much lower (32 %). The small number of shared books between the catalogues of 1846 and 1852 is questionable, but it might be not connected with an actual loss of books from the collection, although some sources

literature were not added as quite different genres of poetry. All the data collected for this study is registered as: https://doi.org/10.5281/zenodo.7846483

confirm that its parts may have been sold to other booksellers ("Knizhnaya kollektsiya…"). Another possible explanation is that some part of the books was afterwards added in a later version of Krasheninnikov's catalogue (1856, not included in the figure).

It is curious that the number of shared books between the Czech library collection and the 19th century catalogues is also very similar and in all the cases (about 30 %) except for Ol'khin catalogue (19 %). This may indicate that some of the books listed in 1846 were neither transferred from Ol'khin to Krasheninnikov's possession, nor recovered in the 20th century: in other words, these books were lost.

The gaps in the overlap between one current library holding and the number of historical catalogues already lead to the question of the rate of loss of the 19th century editions. Smirdin's collection was one example of this process, but more insights can be obtained using the data from multiple library collections.

Contemporary library holdings

The data from the 19th century catalogues served as a starting point for the bibliography of the poetry books issued between 1830 and 1850. During the search of these books in today's library collections, it soon became evident that these catalogues have significant gaps and some collections include books which were never mentioned in either of the four 19th century sources. Thus, the bibliography gradually expanded, mostly drawing from the catalogue of the National library of Russia (NLR). The total number of the poetry books issued between 1830 and 1850 and included in the bibliography resulted in the list of 967 unique titles.

Eleven library collections were used for compiling the bibliography, among them two largest Russian libraries (NLR and Russian State library, RSL), as well as two special collections devoted exclusively to poetry books (Ivan Rozanov's collection of poetry and digitised collection of poetry prepared by the library of the Saint Petersburg State University).[9] The other collections considered are

9 It is to be remarked that the catalogues of some well-stocked libraries as the library of the Pushkin house or rare books collection of the Moscow State University are not digitalised; at the same time, regional Russian libraries possess very low number of books issued before 1917 reflecting the exceptional status of metropolitan libraries.

held outside Russia, as the large Slavonic libraries of the National Library of Finland and that of the National Library of the Czech Republic as well as the Harvard Kilgour collection focused on Russian history and literature (Jackson 13). Number of smaller collections that historically belonged to the western provinces of imperial Russia were also taken into account, since these parts of the empire had their own censorship departments and printing houses and their current holdings may include unique items.

The search in the NLR and RSL started from finding the books listed in the 19th century catalogues and then refined with additional tags —and keyword-based search.[10] For the rest of the library catalogues the whole collection in Russian language dated between 1830 and 1850 was reviewed *de visu*. The list of libraries with the number of poetry books found is presented in Table 1.

Table 1. Number of poetry books found in the contemporary library collections

Abbreviation	Library	N poetry books found (1830-1850)	Approx. collection size (1830-1850)
RU_NLR	National Library of Russia (Saint Petersburg)	924	~50 000[11]
RU_RSL	Russian State Library (Moscow)	529	~20 000
FI_SL	National Library of Finland, Slavonic Library	217	3 500
CZ_SLK	National Library of the Czech Republic, Slavonic Library	197	4 200
US_Harvard	Harvard Library, The Kilgour collection of Russian literature 1750-1920	167	1 800
EE_Ester	United catalogue of Estonian Libraries	58	3 200
LT_VUL	Vilnius University Library	22	4 400
PL_BN	National Library of Poland	19	1 400
PL_BUW	University of Warsaw Library	18	800
RU_Rozanov	Printed catalogue of I.N. Rozanov's collection (Moscow, 1965)	230	–
RU_SPU	Digitised collection "Russian poetry of the 18th and the first third of the 19th century", Saint Petersburg State University	93	–

10 Namely, by examination of the subcatalogues related with poetry as well as by keyword search for books containing the root for "poem / poetry" (стих* / поэт*) or main poetic genres (as an "elegy", "ode", etc.) in any field of bibliographic description.

11 The exact number of books held in NLR and RSL is very roughly estimated, because the catalogues' user interface can frequently output entries which are not unique (e.g., the same book having several catalogue cards) and sometimes beyond the search limits (mainly non-precisely dated books). For the case of NLR, a more rigid total number of books is given in *Table 2*, although it can significantly underestimate the real number of books in the collection at present.

As it is clear from the table, the NLR takes a particular place in preserving the 19th century books, being the legal deposit library[12]. Among other collections, four libraries reach the number of poetry books close to Rozanov's special poetry collection size. Though only about 80 % of the books were sent to the NLR during the first half of the 19th century (Reyser 75-90), it is interesting to see the ratio of poetry books to the total number of books obtained by the library in each year (Table 2). In terms of total numbers, poetry books make up about 3–4 % of all printed editions, so some of the contemporary library collections with a higher ratio are obviously biased towards preserving literary works.

Table 2. Number of poetry books found in NLR in relation to the total number of books according to annual reports reported in (Reyser 75-90)

Year	**1830**	**1831**	**1832**	**1833**	**1834**	**1835**	**1836**	**1837**	**1838**	**1839**	**1840**
Poetry	58	75	72	56	55	44	29	44	43	40	40
% poetry	8.3%	8.7%	9.6%	7.3%	9.3%	6.7%	3.2%	4.3%	3.8%	3.9%	4.2%
Total	692	860	752	764	589	655	897	1021	1103	1016	957

Year	**1841**	**1842**	**1843**	**1844**	**1845**	**1846**	**1847**	**1848**	**1849**	**1850**	**Total**
Poetry	37	42	26	35	37	39	40	41	38	33	924
% poetry	4.1%	4.2%	2.3%	3.2%	3.2%	2.8%	2.7%	1.8%	2.0%	1.9%	3.97%
Total	*901*	*1000*	*1142*	*1079*	*1119*	*1370*	*1464*	*2232*	*1916*	*1737*	*23266*

According to the annual library reports (1830–1850), the number of books deposited to the library doubled towards the end of the 1850s. However, this increase is not visible in the quantity of the poetry books. This can point to two competing reasons: an actual decline in printed poetry, or the lower rates of preservation of 1840s poetry.

To answer the question if the deposit library has a monopolist position keeping books not available anywhere else, Figure 2 displays the distribution of books in the libraries outside NLR in relation to their presence in the 19th century catalogues and the National Library.

12 By the state regulation, all typographies placed in the empire were obliged to send one (1830-1849) or two (1849-1850) copies of all printed books to the library, though this obligation was not always fulfilled (Reyser 75-90).

As it is visible from the flows' sizes, only a small share of poetry books is unique to the NLR. In reverse, a comparable number of books can be found only outside the NLR. Although the 19th century catalogues missed roughly one third of the poetry books found in contemporary libraries, the majority of these books also have at least one copy outside the NLR and only a very small portion of books was lost. Does this fact evidence that the rate of survival for poetry books is very high, despite apparent public indifference to the poetry of the 1830s and 1840s?

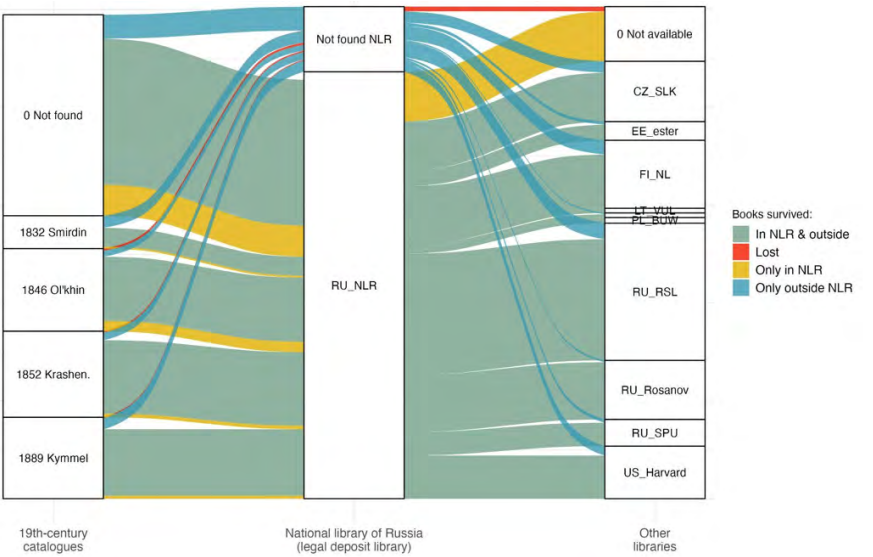

Figure 2. Distribution of poetry books (1830–1850) in other collections in relation to the book presence in the 19th century catalogues and the legal deposit library (NLR)

Poetry books survival rate

Using the compiled bibliography as well as the data on copies distribution in the libraries, it is possible to estimate the survival rate of the books, drawing from the field of population ecology and the problem of biological diversity.

The pioneering attempt to perform an estimation like that was made by L. Egghe and G. Proot who tried to statistically predict the number of lost printed theatre programmes issued in the middle of the 18th century (Egghe

and Proot 257-268). As their model was based on the number of existing copies for each edition and aimed to quantify those which had been lost, the question Egghe and Proot were trying to solve was quickly identified as the unseen species problem and their model criticised for being not as robust as already existing solutions (Burrel 101-105). In biology the estimation of species richness in a particular area is made using so-called abundance data, *i. e.*, the data on the number of encounters of each species in a given sample. While some species are found very frequently, the others have been "seen" only once or twice during the observation period. The least observed species provide data for estimation of the number of even more rare species, which have not been found but, as one suggests, there is still some probability to observe them at least once. Knowing this, Egghe and Proot's estimation of the number of lost or, in other words, unobserved printed documents proves to be "the same context but with time reversed" (Burrell 104).

In the most recent studies on survival of cultural artefacts Mike Kestemont, Folgert Karsdorp, and their colleagues have directly borrowed methods and basic theoretical frame from population ecology to examine the loss rate of mediaeval fictional narratives and manuscripts (Kestemont and Karsdorp 44-55; Kestemont et al. 765-769). As they explain the transition,

> medieval works can be treated as distinct species in ecology, and that the number of extant documents for each work can be regarded as analogous to the number of sightings for an individual species in a sample. Thus, if we treat the available count information for medieval literature as "abundance data", then one can apply unseen species models to estimate the number of lost works in a *corpus* or assemblage. (Kestemont et al. 767)

The concepts of "works" and "documents" in this case oppose immaterial individual narratives and their material mediums (manuscripts) existing in any number of copies. The unseen species model can provide an estimation of the survival rate for both of the categories. As the study revealed, the survival ratio of a mediaeval work (of narrative fiction) is about 68 %, though only about 9 % of the physical documents that contained these texts are estimated to be preserved (Kestemont *et al.* 767).

Since the direct transfer of the method from the natural sciences and its application to the data from humanities may cast doubt on the

obtained results, in their earlier work Kestemont and Karsdorp performed an experiment with simulated data loss. Their results proved that Chao1 delivered the most conservative and robust estimation which is the closest to the ground truth (Kestemont and Karsdorp 50-51). In addition, it is argued that Chao1 estimator brings robust results with little assumptions about data distribution.[13] Thus, we assume that this approach is also feasible for application to the collected data on the poetry books issued between 1830 and 1850. In this case each unique book title will be taken as a species (as a work)[14] and each copy of the work found in the library collections counted as a sighting (a document).

However, it would be unreasonable to calculate the survival rate for all the books together. Five distinct groups of poetry books emerge when dataset is filtered by the content[15] and book volume in the number of pages[16] (Figure 3): 1) short lyric poems either gathered in poetry collections (Lyric-Collections) or 2) issued individually as a small separate editions (Lyric-Booklets); 3) long narrative poems mostly published individually in separate editions (Narrative-Booklets) but 4) sometimes also formed collections (Narrative-Collections); 5) finally, the almanacks and few other collective issues (Almanacks) that unite texts of multiple genres (the books of this group are the only ones in the dataset containing prose).

13 See the description of the assumptions and their influence on the results in the Supplementary materials to (Kestemont *et al.*).

14 For the experiment the second and third editions of the same books were excluded from the dataset as well as 31 books that existed in the 19th century catalogues but not found in any of the observed contemporary collections.

15 See, for example, characteristic titles for books of each category: 1) Lyric-Collections: "Lyrical Poems by Countess Evdokija Rostopchina" [Стихотворения графини Е. Ростопчиной] (1841, 194 pages); 2) Lyric-Booklets: "A Poem on the Temple Consecration in the city of Torzhok on the 15th of September 1842, written by V. N." [Стихотворение по случаю освящения храма в городе Торжке сентября 15-го 1842 года. Соч. В.Н.] (1842, 7 pages); 3) Narrative-Booklets: "The Little Humpbacked Horse: A Russian Fairy Tale by Pyotr Yershov" [Конек-Горбунок: Русская сказка. Соч. П. Ершова] (1834, 122 pages); 4) Narrative-Collections: "Verse Novels by Elisaveta Shahova" [Повести в стихах Елисаветы Шаховой] (1842, 167 pages); 5) Almanacks: "Ladies' Album, compiled from the selected works in Russian poetry" [Дамский альбом, составленный из отборных страниц русской поэзии] (1844, 284 pages).

16 Median number of pages for each group: almanack – 241; collections of lyric poems – 76; collections of narrative poems – 133; separate ed. of lyric poems – 4 (mean: 10); separate ed. of narrative poems – 51. The size in cm is roughly the same for all groups except for usually a smaller format for almanacks.

In general, the distribution of the five types of books confirms scholarly intuitions about this period. Firstly, as expected, the early 1830s are characterised by a large number of almanacks and long narrative forms. Soon after both lost their appeal, most poetry was published as individual poetry collections (Lyric-Collections). The distribution of the number of individually published lyrical poems is less smooth and hints that these editions might be worse preserved than the others. This hypothesis is to be tested statistically using the model of unseen species.

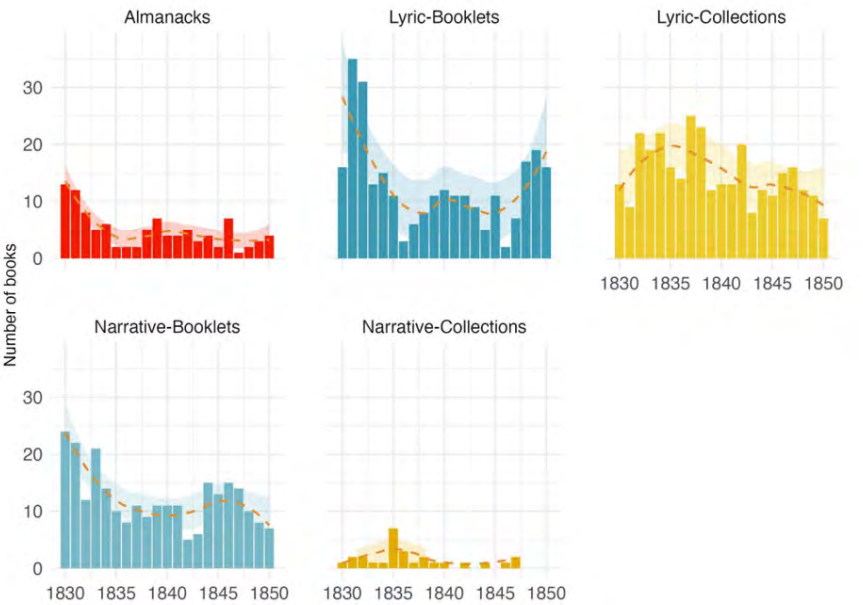

Figure 3. Distribution of poetry books on the timeline

For the estimator Chao1, proposed for use by Kestemont and his colleagues (Kestemont et al. 767), the number of species observed once (singletons) or twice (doubletons) are the most important values. Significantly simplifying the logic behind the estimator explained in (Karsdorp), Chao1 tries to derive the probability of unseen species from the probabilities of singletons and doubletons, assuming that the probability to spot an unseen species will be close to those of singletons and higher than 0. In addition, there is the

"improved" Chao1 estimator that corrects the estimation based not only on singletons and doubletons but also considering the number of species observed three and four times. The latter is important for our data, as shown in Table 3.

Table 3. Abundance data derived from the bibliography of poetry books. Number of occurrences abbreviated as f1 (singletons), f2 (doubletons), f3 (tripletons), and f4 (quadrupletons)

Group	f1	f2	f3	f4	Total works found	Total documents found
Almanacks	6	26	41	16	101	308
Lyric-Collections	64	68	57	54	313	1000
Lyric-Booklets	188	56	15	8	269	387
Narrative-Collections	2	5	2	8	27	108
Narrative-Booklets	69	72	60	31	257	659

After the transformation[17] of the bibliographical data to the counts of sightings, or abundance data, the difference between the five book types becomes more visible. For Lyric-Collections and narrative poems of both types there are many instances of two to three copies for one work to be found, while the Almanacks were preserved even better, in majority of cases having three copies. The distribution of the items observed once (f1) and twice (f2) in our case is drastically different from what was observed for the mediaeval manuscripts, where singletons are much more numerous than works in multiple copies (Kestemont *et al.* 769). This feature underlines the printed nature of the observed collection and questions the applicability of the estimators.

With the abundance data it is possible to estimate survival ratios for each of the groups of the poetry books. As the data contains several groups with majority of tripletons and even some quadrupletons, to take them into account I will mainly discuss the results of the improved *Chao1* (iChao1) estimator (for comparison with other estimators see Table 5). The resulting survival ratios for the works and estimation of the population sizes are shown in Figure 4 and Table 5.

17 For the data transformation and the following analysis and plots I used Python package *copia* made by F. Karsdorp and M. Kestemont, see: https://copia.readthedocs.io/

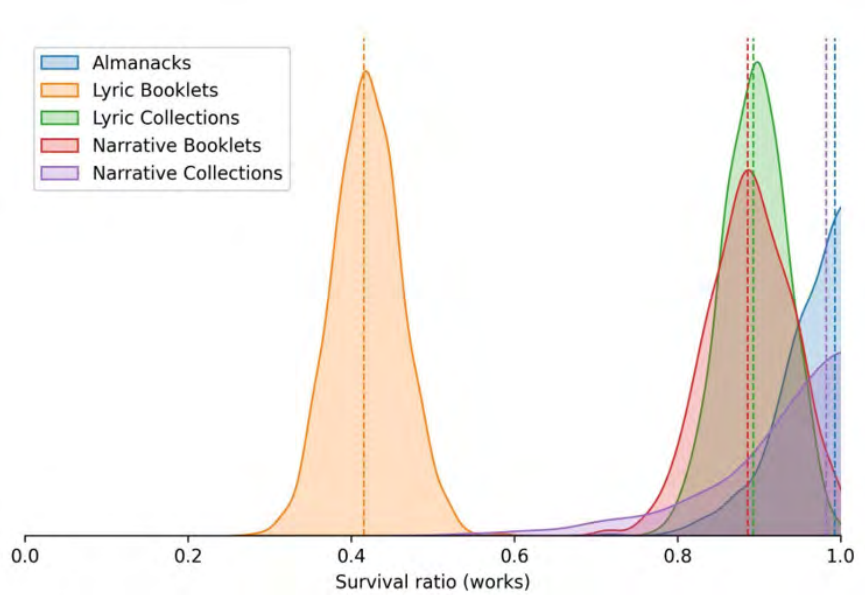

Figure 4. Estimated survival ratios for each of poetry books' groupings

Table 5. Estimated species (works) richness

Group	Works found	iChao1	Chao1	Egghe & Proot
Almanacks	101	101.69	101.69	101.02
Lyric-Collections	313	350.5	343.08	355.49
Narrative-Collections	27	27.48	27.39	27.1
Lyric-Booklets	269	647.27	583.75	597.26
Narrative-Booklets	257	290.01	290.01	293.42

It is visible that for the Almanacks and Narrative-Collections all estimators put the lower bound of species richness at the exact number of already found books. Although there might be simply not enough data for the Narrative-Collections works which are quite rare, this result may confirm that almanack is the best-preserved type of books of this time that include poetry.

For the Lyric-Collections and Narrative-Booklets survival ratio is slightly lower (about 89 %), in other words suggesting that at least 30 individual titles from each group are missed. This result corresponds to the fact that the full bibliography

contains 15 collections of lyric poems and 2 narrative poems that are mentioned in the 19th century catalogues, but not found in the library collections.

The comparison of estimators (Table 5) shows that for the two groups of editions with lyrical poems the improved Chao1 estimator gives less conservative results and predicts a higher number of works than both Chao1 and Egghe & Proot estimators. It is especially clearly seen in the estimated number of Lyric-Booklets. Quite predictably, this group was evaluated as the least preserved type of poetry books with the lower bound richness estimated as at least twice bigger than the number of works that have been found in the bibliography. Figure 5 illustrates the model prediction on the correspondence between the number of found works and documents for Lyric-Collections and Lyric-Booklets groups.

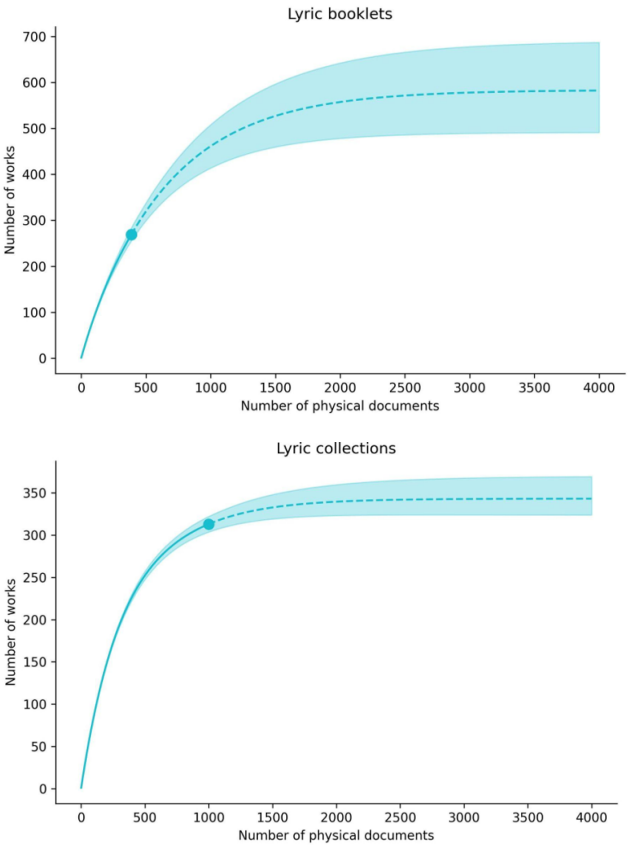

Figure 5. Accumulation curves for Lyric-Collections and Lyric-Booklets

The solid line of the accumulation curves indicates the number of observed works (vertical axis) and documents (horizontal axis) for each group corresponding to the abundance data (Table 3). The dashed line shows the model estimation on how many physical copies are to be observed to find books that have not been found yet. This can be read as: it would take about five hundred more lyrical collections to be found in the library catalogues (quite unrealistic task) to have a probability to find about 30 "unobserved" ones. In the case of Lyric-Booklet, the same amount of observed new documents would likely double the bibliography of this type of editions.

Conclusion

The path of this small-scope bibliography of poetry books shows how different sources and methods can possibly bring new knowledge on the features of a particular sample of literary works. In terms of bibliography, the cases of Smirdin's library and distribution of copies in contemporary libraries are not considered here as exhaustive. Obviously, more contemporary collections can be taken into account as well as additional data added for the better tracing of books from Smirdin's library.

However, as I tried to show, computational methods can be used to assess the incompleteness of samples when collecting data on printed sources. In our case the estimation of survival ratios demonstrates that data contains completely different types of poetry books that survived at different rates. In general, the results provided by the model are consistent with evidence from the literary history of this period. Thus, the application of the unseen species model seems to be an appropriate way to examine the survivorship bias in a particular collection.

Some notes still can be added regarding the use of conservative estimators for not-that-rare species as multi-copied books. It is predictable that the model estimates almost perfect survivability for almanacks (with the majority of tripletons), however, the assessment of the booklets is the most insightful. Considering the size of the dataset and assuming that for most books print run size should have been larger than few dozens of copies, it was realistic to use sampling with replacement, assuming that the appearance of one copy

in a collection does not influence its presence in another one.[18] However, this might not be true for the small booklets as this print form was usually occasional and could be published in scarce copies. As the estimation of the number of lost booklets seems one of the main findings of this study, it is to be verified if the resulting estimation was not impacted by the sampling issue.

Nevertheless, the mere possibility to estimate the original population size seems to be missing part of the bias correction in computational literary research. Returning to the question of differences between existing *corpus* of Russian poetry and the "exhaustive" one assessed in this paper, it should be concluded that in this case the selection of secondary sources makes the two assemblages almost incomparable. At the same time, it is certain that quite adequately preserved poetry collections of the 1830–1850 would give a lot of data for compiling "the unread"part of the *corpus* presenting a large fraction of the true literary population. Moreover, the estimation suggests that the loss of poetic texts most likely happened in small occasional booklets, a bit less relevant in terms of *corpus* building. Finally, obtaining these conclusions from quantifiable data seems to be more valuable than just relying on intuitions or tradition.

Works cited

Balakin, Aleksey Yu. "Poeticheskie Knigi 1834-1850". *Livejournal*, 9 June 2010. Web. Accessed 15 Jan. 2023.

Bukhshtab, Boris Ya. "Russkaya Poeziya 1840-kh – 1850-kh Godov". *Poety 1840-1850-kh Godov*. Leningrad, Sovetskiy Pisatel', 1972, pages 5-60.

Burrell, Quentin L. "Some comments on 'The estimation of lost multi-copy documents: A new type of Informetrics Theory' by Egghe and Proot". *Journal of Informetrics*, vol. 2, no. 1, 2008, pages 101-105. DOI: https://doi.org/10.1016/j.joi.2007.07.002

Egghe, Leo, and Goran Proot. "The estimation of the number of lost multi-copy documents: A new type of Informetrics Theory". *Journal of Informetrics*, vol. 1, no. 4, 2007, pages 257-268. DOI: https://doi.org/10.1016/j.joi.2007.02.003

Huber, Alexander, ed. *Eighteenth-Century Poetry Archive*, 17 Jan. 2023. Web. Accessed 17 January 2023.

18 On this problem in connection with unseen species models, see (Wevers et al. 189-197).

Jackson, William A. Preface. *Kilgour Collection of Russian Literature, 1750-1920*. Cambridge, Massachusetts, Harvard College Library, 1959.

Kestemont, Mike, and Folgert Karsdorp. "Estimating the loss of Medieval Literature with an Unseen Species Model from Ecodiversity." *CEUR Workshop Proceedings*, vol. 2723, 2020, pp. 44-55.

Kestemont, Mike, et al. "Forgotten books: The application of Unseen Species Models to the Survival of Culture." *Science*, vol. 375, no. 6582, 2022, pages 765-769. DOI: https://doi.org/10.1126/science.abl7655.

Kishkin, Lev S."Knizhnoe sobranie A. F. Smirdina v Prage". *Vremennik Pushkinskoy Komissii, 1974*. Leningrad, Nauka, 1977, pages 148-155.

Knizhnaya Kollektsiya A. F. Smirdina v Fondakh Gosudarstvennoy Publichnoy Istoricheskoy Biblioteki. Accessed 15 Jan. 2023.

Koblents, Ioel' N. "«Rospis'» biblioteki A.F. Smirdina (Ee znachenie v istorii i statistike pechati pushkinskoy pory)". *Kniga. Issledovaniya i Materialy*, vol. 26, Moskva, Kniga, 1973, pages 80-93.

Korchagin, Kirill M. "Poeziya XX veka v Poeticheskom podkorpuse Natsional'nogo Korpusa Russkogo Yazyka: Problema reprezentativnosti". *Trudy Instituta Russkogo Yazyka Im. V.V. Vinogradova*, no. 6, 2015, pages 235-56.

Levin, Grigoriy L. "Bibliograficheskiy repertuar russkoy knigi 1826–1917 gg.: Istochniki i puti formirovaniya". *Bibliotekovedenie*, vol. 66, no. 2, 2017, pages 147-153.

Métrique En Ligne: Corpus Poétiques, Relevés Métriques et Outils d'analyse. June 2022. Web. Accessed 15 Jan. 2023.

Orlov, Vladimir N."Drugie poety dvadtsatykh – tridtsatykh godov". *Istoriya Russkoy Literatury*, vol. 6, AN SSSR, 1953, pages 478-498.

Reyser, Solomon A."Ob istochnikakh russkoy knizhnoy statistiki". *Sovetskaya Bibliografiya*, vol. 1 no. 20, 1946, pages 75-90.

Rose, Jonathan, and Simon Eliot. *Companion to the History of the Book*. Wiley Blackwell, 2019.

Rospis' Rossiyskim Knigam dlya Chteniya iz Biblioteki Aleksandra Smirdina, sistematicheskim Poryadkom Raspolozhennaya. Saint Petersburg, A. Smirdin, 1828.

Ruiz Fabo, Pablo, and Helena Bermúdez Sabel, eds. *DISCO: Diachronic Spanish Sonnet Corpus*. v5.0, Zenodo, 24 Feb. 2023. DOI: https://doi.org/10.5281/ZENODO.7675512.

"Russkaya Literatura". *Severnaya Pchela*, no. 198, 3 Sept. 1838, page 791.

Savitskiy, Ivan. "Novaya zhizn' russkoy biblioteki v Prage: O sud'be knizhnogo sobraniya A.F. Smirdina". *Sovetskoe Slavyanovedenie*, no. 4, 1967, pages 96-98.

Smirnov-Sokol'skiy, Nikolay P. *Russkie Literaturnye Al'manakhi i Sborniki XVIII–XIX vv*. Moskva, Kniga, 1965.

Vatsuro, Vadim E. "Poeziya 1830-kh Godov". *Istoriya Russkoy Literatury*. Vol. 2. Leningrad, Nauka, 1981, pages 362-379.

Wevers, Melvin, et al. "What shall we do with the Unseen Sailor? Estimating the size of the Dutch East India Company using an Unseen Species Model." *CEUR Workshop Proceedings*, vol. 3290, 2022, pages 189-97.

About the author

Antonina Martynenko, MA in Russian and Slavonic Philology, currently a PhD candidate and a Junior Research Fellow in Russian and Slavonic Philology at the University of Tartu.