**Создание модели для автоматического распознавания русскоязычных печатных текстов XVIII века с помощью Transkribus**

**Авторы:** Кузнецова Алёна, Староверова Полина

**Аннотация:** В статье рассказывается о модели для распознавания русскоязычных печатных текстов XVIII века, созданной на платформе Transkribus. Помимо описания самой модели, в статье содержится информация об особенностях печатных текстов этого периода и типографиях, существовавших в то время.

**Ключевые слова:** Транскрибус, Transkribus, машинное распознавание текста, Петровский гражданский шрифт, типография Академии Наук, типография Московского университета

**Оглавление**

[Введение 2](#_Toc107439078)

[Контекст печати в XVIII веке в России 4](#_Toc107439079)

[Процесс обработки текстов в Transkribus 7](#_Toc107439080)

[Наша модель 9](#_Toc107439081)

[Выводы и направления дальнейшей работы 11](#_Toc107439082)

[Список литературы 12](#_Toc107439083)

[Приложение 1 13](#_Toc107439084)

[Приложение 2 14](#_Toc107439085)

# 

# **Введение**

Русскоязычные печатные тексты XVIII века – это большой и разнообразный корпус текстов, важный для исследователей из разных сфер. Помимо художественной литературы (как отечественной, так и переводной), в это время печатались книги, посвященные точным наукам (географии, архитектуре, инженерному делу, биологии и пр.), исторические очерки, путевые заметки, первые российские журналы, словари, разнообразные наставления, указы и многое другое.

Издания этого времени в последние годы оцифровываются в библиотеках и в рамках проектов, посвященных сохранению книжных памятников. Например, поиск текстов за период с 1700 по 1800 год на Google books[[1]](#footnote-1) выдает примерно 15000 текстов, а поиск на НЭБ[[2]](#footnote-2) 4000 текстов. Однако в случае с Google books мы не можем быть уверены, что все эти тексты русскоязычные. К сожалению, хотя в Google books выставляется фильтр по языку издания, в поисковую строку все равно приходится вводить какое-либо слово. Сервис уже сам подбирает тексты, в которых содержится указанное слово, входящие в указанный в фильтре период издания. При этом встроенная в Google books автоматическая модель распознавания текста ошибается настолько сильно, что «находит» в латиноязычных или немецкоязычных книгах слова, написанные кириллицей. Тем более не стоит ожидать от этой модели качественного распознавания букв, чье написание в тексте определенного периода значительно отличается от более часто встречающегося в печати вида буквы.

Другие платформы, хранящие книжные памятники этого периода, как например, уже упомянутый НЭБ, к сожалению, не оснащены встроенным инструментом распознавания текста. Качественно распознанных и приведенных к формату, пригодному для обработки цифровыми методами, текстов XVIII века очень мало, что значительно затрудняет изучение этих текстов исследователями. Наличие инструмента, позволяющего быстро распознать текст из оцифрованной книги, могло бы привлечь больше исследователей к изучению русскоязычных печатных текстов XVIII века.

Так появилась идея о создании модели для автоматического распознавания русскоязычных печатных текстов XVIII века. В качестве инструмента для создания этой модели был выбран Transkribus – комплексная платформа, предназначенная для изучения и распознавания рукописных и печатных текстов с помощью искусственного интеллекта для их последующей расшифровки в текст, читаемый компьютером. Помимо распознавания текста и устранения шумов[[3]](#footnote-3), платформа умеет распознавать изображения, орнаменты, химические формулы, нотную запись и другие элементы страниц, что является дополнительным преимуществом для изучения исторических материалов, в которых зачастую встречаются узоры и сложные по устройству инициалы первых абзацев для украшения страниц (прил. 2, рис. 1)[[4]](#footnote-4).

Преимущество этой программы в том, что, в отличие от таких аналогов распознавания данных документа как Abbyy FineReader, OCR Cunei Form или Readiris Pro, которые умеют только находить текст, Transkribus предоставляет возможность не просто самим обучить модель, но и поделиться ей с другими пользователями, добавив в коллекцию уже загруженных в программу моделей. Это означает, что клиенты программы могут работать не только локально, но и обмениваться полученными результатами для развития дальнейших исследований, что, в свою очередь, существенно сокращает время распознавания материала.

Говоря об обмене опытом при помощи Transkribus’а, стоит упомянуть об уже существующих моделях[[5]](#footnote-5), созданных именно для печатных текстов. Так, например, командой самой платформы (Transkribus Team) были созданы модели расшифровки нескольких языков одновременно, а именно: голландского, английского, финского, французского, немецкого и шведского. Из доступных к использованию на данный момент 94 моделей, только 29 могут распознавать печатный текст, из которых нет ни одной для русского языка. Следовательно, цель нашей работы заключается в том, чтобы создать и предоставить доступ к первой модели анализа и расшифровки русскоязычных печатных материалов на базе Петровского Гражданского шрифта и типографий Академии Наук и Московского университета.

# **Контекст печати в XVIII веке в России[[6]](#footnote-6)**

*Первопечатные книги (XVI-XVII вв.)*

Книгопечатание в России появилось в конце 50-х – в начале 60-х годов XVI века при Иване Грозном. Первые книгопечатные книги по содержанию делились на три группы:

1) церковно-служебные;

2) предназначенные для чтения и изучения вне церкви (жития, нравоучительные статьи, иллюстрированные Апостол и Евангелие для чтения дома);

3) книги светского назначения (буквари, «Грамматика» Мелетия Смотрицкого, «Обед душевный» и «Вечеря душевная» Симеона Полоцкого и Симона Ушакова).

Характерной особенностью первопечатных книг было обилие надстрочных знаков: нужных либо для сокращения текста, либо для обозначения ударений и придыханий (важных для чтения вслух в церкви). Также в первопечатных книгах широко применялись разные начертания букв, например, широкое и узкое *о*, *о* с точкой либо двумя точками внутри.

*Печать XVIII в. (при Петре I)*

При Петре I в России впервые широко организовалось печатание гражданских книг: математических, военных, по градостроительству, географических, исторических, навигационных и по другим направлениям. Производство русскоязычных книг было в Голландии в типографиях Яна Тессинга и Ильи Федоровича Копиевского, получивших разрешительную грамоту от государя.

В 1708 году Петр I утвердил стандарт нового гражданского шрифта, корректуры в который вносились до 1710 года. Проведенная реформа была крайне важна: изменился состав азбуки и одновременно с этим изменилась графика шрифта. Важно учитывать то, что при разработке гражданской азбуки были исключены знаки ударений, в том числе придыханий и сокращений, а также введены арабские цифры вместо обозначения чисел буквами, которые затрудняли арифметические действия.

При этом многие «лишние» буквы все равно сохранились (например, *Ѯ, Ψ, ω*), так как их упразднению препятствовала церковь.

*Печать XVIII в. (после Петра I)*

По данным каталога русской книги гражданской печати XVIII века[[7]](#footnote-7), в Петербурге и Москве в конце XVIII века насчитывалось до 50 типографий, из которых около 40 были частными. Однако в сороковых годах XVIII века графика русского типографского шрифта определялась в основном типографией Академии наук, основанной в 1727 году в Петербурге, в дальнейшем также типографией Московского университета, основанной в 1756 году.

В связи с этим развитие русского типографского шрифта во второй половине XVIII века рассматривается в трех направлениях:

1) шрифты петербургской типографии Академии наук;

2) шрифты типографии Московского университета;

3) шрифты других государственных и частных типографий Москвы и Петербурга.

*Типография Академии наук*

Начиная с 1730-х годов в печати начинают появляться другие шрифты, отличающиеся видом от того, который появился при Петре I. В 1733 году появился шрифт антиква (рис. 2). Новый шрифт отличался от предыдущих уменьшенным междустрочным пробелом в результате сокращения выносных элементов. По сравнению с петровским шрифтом среднего размера, шрифт антиква несколько уже. Нижние завитки букв *ц* и *щ* в прописном начертании заменены обычным штрихом влево.

К середине XVIII века было уже 12 кириллических шрифтов: каждый в разных масштабах, для некоторых были еще и курсивы, кроме того, были отлиты литеры для латинского, греческого и немецкого языков. Они были собраны в издании образцов шрифтов под названием «Пробная книга всем азбукам, знакам и типографским украшениям, которыя при императорской Академии наук в типографии находятся» (было выпущено в 1748 году и отразило новый этап в развитии русского гражданского шрифта) (рис. 3). Оно было переиздано в 1788 году и отразило произошедшие стилистические изменения в существующих шрифтах.

По данным Абрама Шицгала[[8]](#footnote-8), самыми распространенными шрифтами для набора текста со второй половины XVIII века были три вида шрифтов: миттель антиква, терция антиква и гробе цицеро. Между этими шрифтами есть разделение - использование конкретного шрифта часто зависело от тематики произведения, поэтому иногда в рамках одной книги могли использоваться разные шрифты.

*Типография Московского университета (Новиков с 1779 по 1789 гг.)*

В типографии Московского университета отливались свои шрифты (рис. 4), главные отличия которых от шрифта петербуржской типографии заключались в следующем:

- меньший контраст между основными и соединительными штрихами;

- более строгое начертание контура букв, освобожденное, в частности, от типичных древнерусских начертаний;

- более строгое построение буквы *к*;

- уменьшение размера треугольных окончаний в буквах *е*, *т*, *д*.

*Частные типографии*

В основном частные типографии использовали либо петербургский, либо московские литеры. Однако есть два исключения:

1) типографии, которые перерабатывали московские и петербургские шрифты: в Петербурге это Императорская типография (рис. 5), типографии Морского кадетского корпуса, Вейтбрехта и Шнора; в Москве это типография Селивановского и товарища;

2) типографии, которые заказывали литеры в Париже: небольшая петербургская типография Корпуса чужестранных единоверцев и частная типография И. К. Шнора (рис. 6).

# **Процесс обработки текстов в Transkribus**

Transkribus является первой междисциплинарной платформой, которая предоставляет специалистам в гуманитарных науках возможность обрабатывать не только средневековые и возрожденческие тексты, написанные относительно разборчивым минускулом, но и рукописные материалы любого типа[[9]](#footnote-9). Что касается автоматической транскрипции, для европейских языков доступно большое количество готовых данных, однако постепенно хорошие результаты достигаются и для других языков, в том числе славянских[[10]](#footnote-10). Речь идет в первую очередь о рукописных текстах, число моделей для печатных текстов крайней невелико даже для европейских текстов.

Как уже утверждалось ранее, Transkribus включает инструменты для осуществления всего процесса распознавания текста, то есть для выполнения расшифровки материала не требуется никаких дополнительных инструментов. Процесс подготовки извлечения текстового контента состоит из следующих шагов:

1) анализ изображения;

2) удаление шума из изображения;

3) выравнивание текста;

4) анализ макета;

5) распознавание;

6) преобразование текста изображения в машинно-кодированный текст.

В части предварительной обработки первым важным шагом является бинаризация изображения. Цветное или полутоновое изображение необходимо преобразовать в черно-белое, чтобы текст был максимально изолирован[[11]](#footnote-11). Этот процесс чрезвычайно важен для следующего шага — удаления шума из изображения. С точки зрения анализа изображений исторических рукописей, шум может быть любым знаком, который не является частью основного текста, например, ненужные точки, проступание чернил с оборотной стороны, грязь или другие виды помех. После того, как текстовая часть изображения была найдена, ее необходимо выровнять. Обычно отсканированные изображения уже достаточно выровнены; однако процесс устранения перекоса приведет к тому, что текст строки будет располагаться в еще более горизонтальной плоскости[[12]](#footnote-12). Анализ макета текста и его последующее распознавание начинается после сегментации страницы к изображению: в этот момент обнаруживаются различные структурные элементы, а именно: абзацы, примечания, не являющиеся частью основного текста, таблицы, рисунки и так далее. Последним шагом в процессе распознавания материала является преобразование текста изображения в машинно-кодированный. К сожалению, достичь полной автоматической транскрипции пока невозможно, поэтому обычно для улучшения результата применяется ручная коррекция текста. Если CER (Character Error Rate)[[13]](#footnote-13) - частота ошибок в символах - недостаточно низок, время, затраченное на процесс исправления, вероятно, будет таким же, как и время, необходимое для ручной транскрипции[[14]](#footnote-14).

После извлечения текстового материала можно приступать к созданию модели. Согласно руководству Transkribus, для начала обучения модели требуется 25–75 расшифрованных страниц. Чем больше видов шрифтов в корпусе для обучения, тем большее количество страниц нужно использовать для создания достоверной модели. Transkribus также предоставляет возможность добавления уже готовой базовой модели при обучения новой того же стиля письма. Это существенно облегчает процесс обучения и повышает точность модели[[15]](#footnote-15).

# **Наша модель**

За время работы над проектом было создано три модели: 1) модель, обученная на текстах, напечатанных в типографии Академии Наук (далее *модель 1*); 2) модель, обученная на текстах типографии Московского Университета (далее *модель 2*) и 3) финальная модель, обученная на текстах типографии Академии Наук, Московского Университета, типографии Вильковского и Галченкого и текстах, напечатанных Петровским Гражданским шрифтом (далее *модель 3*).

*Модель 1* была обучена на небольшом корпусе (6459 слов, 1154 строк) из книг, напечатанных в типографии Академии Наук, так как несколько вручную распознанных книг этой типографии предоставил нам НКРЯ[[16]](#footnote-16). По итогам тестирования этой модели уже вскрылись некоторые проблемы. Во-первых, модель, обученная на страницах, где размечен только текст и ничего больше, воспринимает как текст и элементы страницы, на которых текста нет: «видит» буквы в орнаментах, картинках и технических повреждениях. Для того, чтобы этого избежать, необходимо размечать с помощью инструментов Transkribus’а все элементы страницы. Одной из особенностей печати в XVIII веке является обильное использование заглавного и строчного шрифта вперемешку. Модель плохо справлялась с тем, чтобы отличать друг от друга заглавные и строчные буквы (рис. 7).

При этом *модель 1* практически не допускала ошибок в различении похожих в написании букв (за исключением путаницы между ѣ и *ъ*). Однако относительно успешные результаты модель показывала только на текстах типографии Академии Наук. Книги, напечатанные в типографии Московского университета или тексты, набранные Гражданским шрифтом, модель распознавала плохо. Шицгал писал о том, что уже начиная с 60-х годов XVIII века начали накапливаться различия между петербургскими и московскими шрифтами, а уже к 80-м шрифты резко отличались друг от друга[[17]](#footnote-17). Поэтому изначально предполагалось, что нам потребуется минимум три модели для текстов этого периода: модель для Петровского гражданского шрифта; модель для типографии Академии Наук и типографий, заказывающих шрифты у нее и модель для типографии Московского университета и типографий, заказывающих шрифты у этой типографии.

Именно поэтому *модель 2* была обучена на корпусе текстов, напечатанных в типографии Московского университета (4783 слов, 1039 строк). Эта модель также, как и предыдущая, плохо работала на текстах, набранных Петровским гражданским шрифтом, однако очень хорошо показала себя на текстах типографии Академии Наук. Качество распознавания букв было почти таким же высоким, как на текстах типографии Московского университета. Из чего мы сделали вывод, что различия между шрифтами этих двух типографий не настолько критичны, чтобы существовала необходимость в двух разных моделях и решили, что можно попробовать обучить единую модель. Однако *модель 2*, как и *модель 1*, «видела» буквы в элементах служащих для украшения страницы (рис. 8).

Для решения этой проблемы в мы решили поэкспериментировать с элементами структурной разметки (тэгами). Это встроенный в Transkribus инструмент, позволяющий обучить модель таким образом, чтобы схожим элементам страницы она могла присваивать одинаковые названия, если определенный тэг в корпусе выделен более 50 раз. Мы ограничились шестью различными тэгами: 1) Page-number – номер страницы; 2) Marginalia – заметка на полях; 3) Book decoration – украшение; 4) Map – карта; 5) Header – колонтитул; 6) Initial – буквица. К сожалению, в *модель 3* структурная разметка так и не была включена, но обязательно будет добавлена в модель, предназначенную для публичного доступа.

*Модель 3*, обученная на текстах типографии Академии Наук, Московского Университета, типографии Вильковского и Галченкого и текстах, напечатанных Петровским Гражданским шрифтом, была создана на корпусе, в три раза превосходящем объем корпуса, на котором создавалась *модель 1* (22432 слов, 4445 строк). Перечень текстов, на которых она была обучена, можно найти в Приложении 1. Качество модели в Transkribus’е определяется показателем CER на проверочном корпусе (Validation Set)[[18]](#footnote-18). У нашей модели CER на проверочном корпусе равен 3.64% (проверочный корпус составил 10 % от размера корпуса, на котором модель обучалась).

*Модель 3* была протестирована на текстах, которые не входили в состав корпуса для ее обучения. Результаты тестирования на текстах, набранных Петровским гражданским шрифтом, показали, что модели удается хорошо распознавать отчетливо набранный текст. Она не путает похожие в написании буквы *ш* и *т; ъ* иѣ (рис. 9). При этом она ошибается в распознавании заглавных и строчных букв и не распознает как единый символ инициал, который значительно выбивается из размеров строки (рис. 10).

Для тестирования распознавания текстов типографии Московского университета была намеренно выбрана та же страница, что и для одного из тестирований *модели 2* (сравните рис. 8 и рис. 11). В отсутствии внедренной структурной разметки модель еще тщательнее пытается распознать буквы в орнаменте, однако абсолютно верно распознает номер страницы и слово «сентября» (*свитибри* у более ранней модели).

Очень интересная ситуация сложилась с распознаванием текста типографии Академии Наук. Сначала для тестирования был выбран курсивный шрифт, характерной особенностью которого у этой типографии является начертание буквы *в*, близкое к скорописи, похожее на букву *п*. Модель трижды правильно распознала букву *в* (даже в случае, когда буквы *п* и *в* были почти рядом – слово *купивши*) и ошиблась в слове *содержавшимся* (рис. 12).

При этом текст, набранный не курсивным шрифтом, модель распознала верно (рис. 13). На этой же странице модель абсолютно корректно выделила примечание на полях в отдельную область текста, но по какой-то причине распознала в слове *л*ѣ*т* греческие буквы, которые в корпусе, на котором модель обучалась, встречались только один раз (рис. 14).

# **Выводы и направления дальнейшей работы**

Результаты тестирования *модели 3* уже показывают высокое качество ее работы на текстах XVIII века и сравнительно небольшое количество ошибок. Однако, на наш взгляд, перед тем как выложить ее в открытый доступ для всех пользователей Transkribus’а, модель необходимо усовершенствовать. Во-первых, об этом уже шла речь ранее, в модель нужно добавить структурную разметку, тем более что данные, на которых ее можно обучить, уже размечены и подготовлены. Структурная разметка может «научить» модель не искать буквы в рисунках и орнаментах, а также выделять инициалы в отдельный элемент. Во-вторых, корпус необходимо дополнить б**о**льшим количеством размеченных страниц с курсивными шрифтами и таблицами. Распознавание такого текста нашей моделью на данный момент самое низкое по качеству. В-третьих, следует протестировать модель на текстах тех типографий, которые не входили в состав корпуса, на котором она обучалась. К сожалению, в процессе работы выяснилось, что книг, отпечатанных в небольших частных типографиях, оцифровано очень мало, но мы надеемся найти достаточное количество текстов. И, если качество распознавания текста будет низким, разметить эти тексты и добавить в корпус для обучения модели.

После доработки модели мы планируем сделать ее общедоступной для пользователей Transkribus, а также опубликовать необходимые инструкции о том, как использовать Transkribus и нашу модель в частности, на русском языке для того, чтобы больше ученых могли с помощью этого инструмента проводить свои исследования.

# **Список литературы**

1. Milioni, N. *Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars. Department of ALM. Theses within Digital Humanities*, 2020.
2. Muehlberger, Guenter, et al. "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study." Journal of documentation (2019) URL: https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html (Дата обращения: 12 июня 2022 года).
3. Сводный каталог русской книги гражданской печати XVIII века 1725-1800: В 5 т. Т. V. Указатели. М.: Книга, 1967.
4. Шицгал А. Г. Русский типографский шрифт: вопросы истории и практика применения. М.: Книга, 1985.

# **Приложение 1**

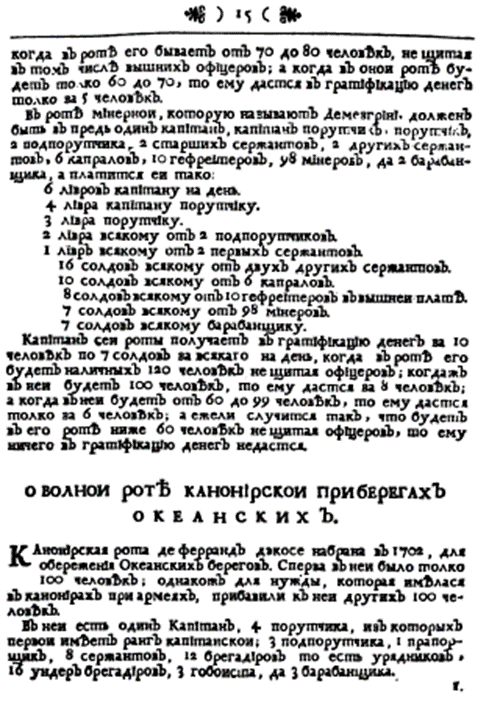
**Список текстов, вошедших в корпус для обучения финальной модели**

1. Архитектура воинская гипотетическая и эклектическая, 1709, Московский печатный двор.
2. Дневные записки путешествия доктора и Академии Наук адъюнкта Ивана Лепехина по разным провинциям Российского государства 1768 и 1769 году. Часть 1, 1795, типография Академии Наук.
3. Ежемесячные сочинения к пользе и увеселению служащие, 1755, типография Академии Наук.
4. Журнал или дневные записки путешествия Капитана Рычкова по разным провинциям Российского государства в 1769 и 1770 году, 1795, типография Академии Наук.
5. Известие о дворянах российских, 1790.
6. Истинная политика знатных и благородных особ [перевод книги Н. Ремона де Кура с французского], 1745, типография При Императорской Академіи Наукъ.
7. Историческое изображение жизни и всех дел славного женевца, Франца Яковлевича Лефорта, 1800, типография Московского университета.
8. История о странствиях вообще по всем краям земного круга. Часть 1, 1782, типография Московского университета.
9. Краткое Математическое изъяснение землемерия межевого 1757 года, 1757.
10. Краткое описание всех случаев касающихся до Азова от создания сего города до возвращения оного под Российскую державу, 1782, типография Академии Наук.
11. Краткое описание о воинах из книг Цезариевых, 1711, Московский печатный двор.
12. Кто может быть добрым гражданином и верным подданным, 1796, типография Московского университета.
13. Московский Журнал. Часть 1. Январь-март 1791, 1791, типография Московского университета.
14. Непостоянная фортуна, или похождение Мирамонда. Часть 2, 1781, типография Московского университета.
15. Несчастные приключения Василья Баранщикова мещанина Нижнего Новгорода в трех частях света: в Америке, Азии и Европе с 1780 по 1787 год, 1787, типография Вильковского и Галченкого.
16. Несчастный француз, или жизнь кавалера Беликурта, описанная им самим, 1764, типография Московского университета.
17. Новое голландское корабельное строение, 1709.
18. О должности человека и гражданина по закону естественному. Книги первая и вторая, 1726.
19. Описание лагеря, собранного под высочайшею Ея Императорского Величества собственною командою при Красном селе, 1765, типография Академии Наук.
20. Описание моровой язвы, бывшей в столичном городе Москве с 1770 по 1772 год, 1775, типография Московского университета.
21. Опыт о пользе трезвой жизни, 1768, типография Московского университета.
22. Путешественные записки Василья Зуева от С. Петербурга до Херсона в 1781 и 1782 году, 1787, типография Академии Наук.

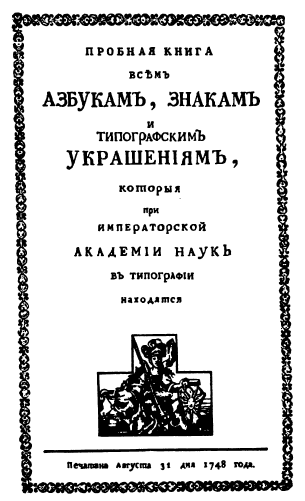
# **Приложение 2**



*Рисунок 1.* Из книги «Истинная политика знатных и благородных особ»



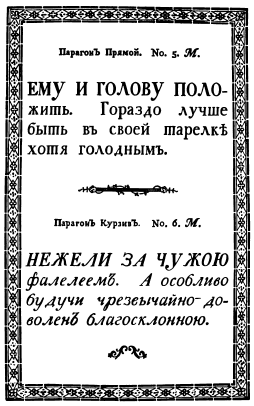
*Рисунок 2*. Из «Русского типографского шрифта» А. Г. Шицгала, 1985. С. 54.



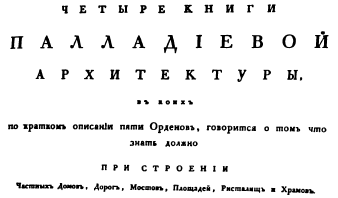
*Рисунок 3*. Из «Русского типографского шрифта» А. Г. Шицгала, 1985. С. 56.



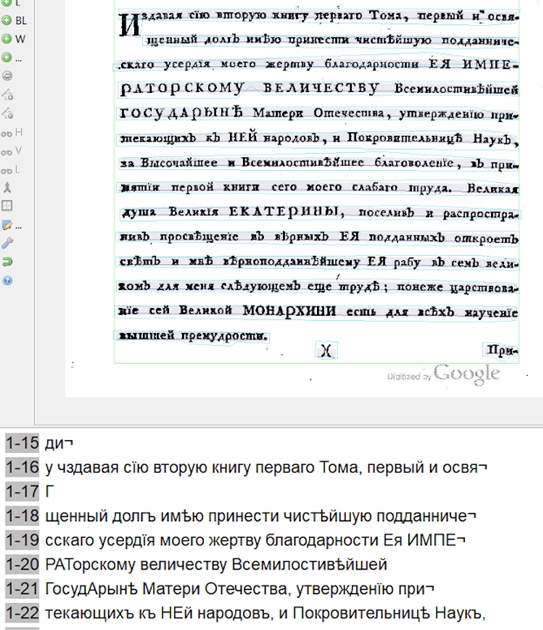
*Рисунок 4*. Из «Русского типографского шрифта» А. Г. Шицгала, 1985. С. 78.



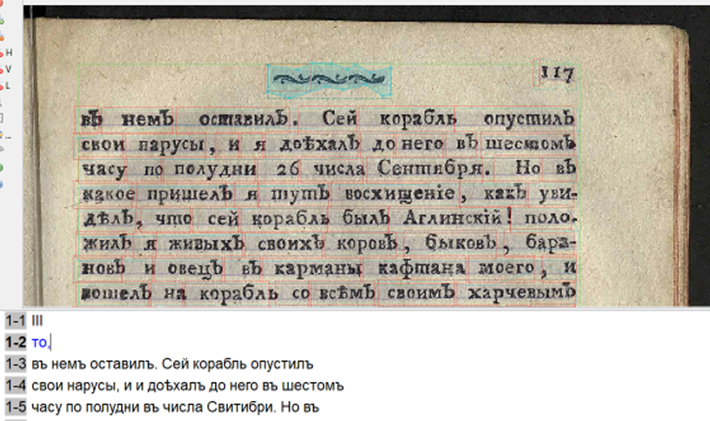
*Рисунок 5*. Из «Русского типографского шрифта» А. Г. Шицгала, 1985. С. 90.



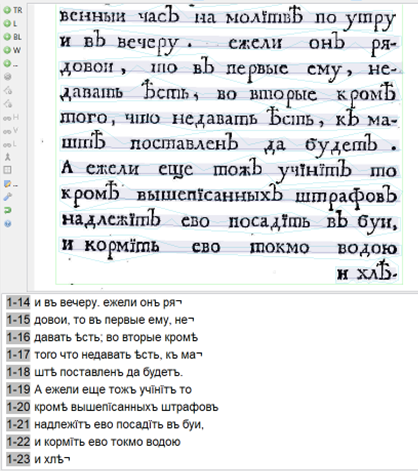
*Рисунок 6*. Из «Русского типографского шрифта» А. Г. Шицгала, 1985. С. 96.



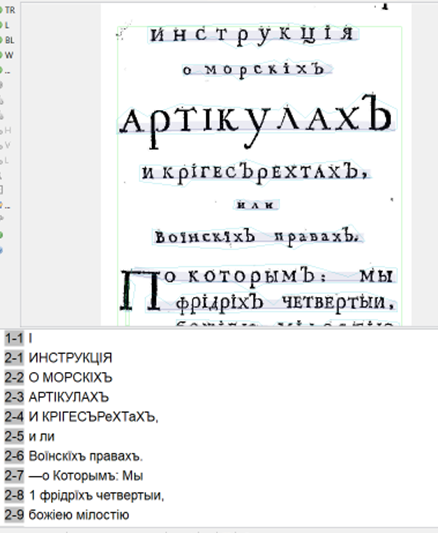
*Рисунок 7*. Из «Исторического описания российской коммерции», 1782.



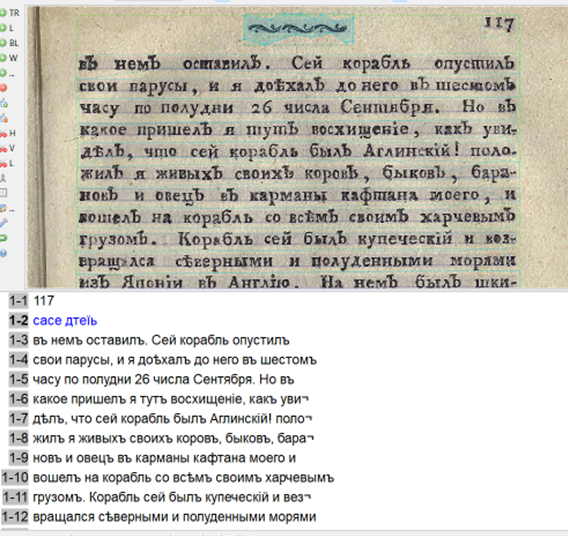
*Рисунок 8*. Из «Путешествий Гулливеровых. Книга 1», 1780.



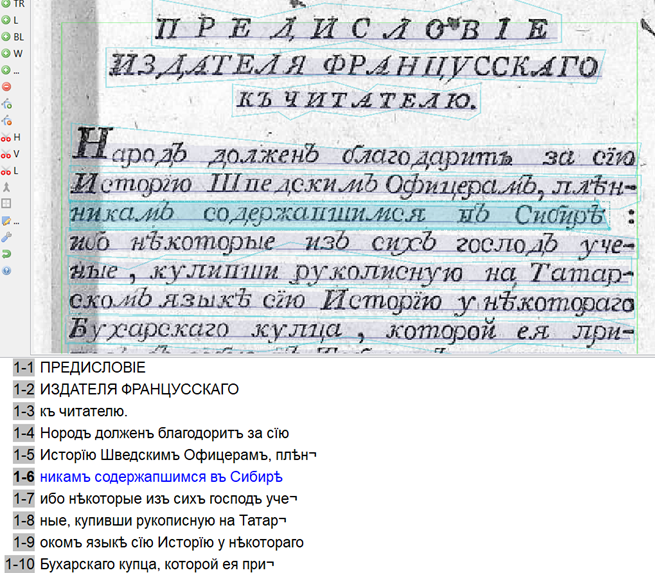
*Рисунок 9*. Из «Инструкции о морских артикулах», 1735.

**

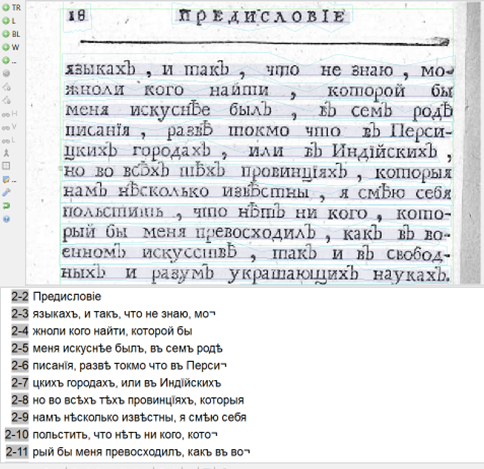
*Рисунок 10*. Из «Инструкции о морских артикулах», 1735.



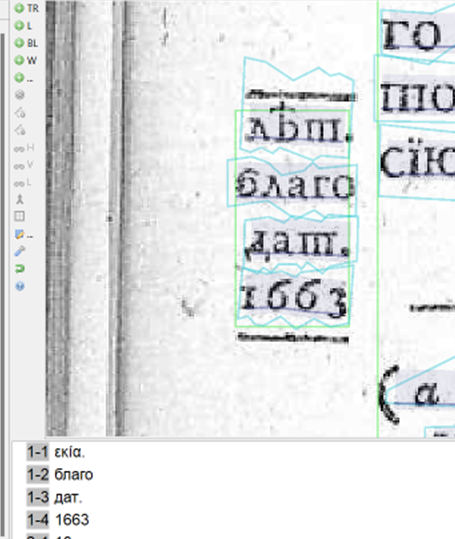
*Рисунок 11*. Из «Путешествий Гулливеровых. Книга 1», 1780.



*Рисунок 12*. Из «Родословной истории о татарах», 1768.



*Рисунок 13*. Из «Родословной истории о татарах», 1768.



*Рисунок 14*. Из «Родословной истории о татарах», 1768.

1. Google books. URL: [https://books.google.ru](https://books.google.ru/) (Дата обращения: 07 июня 2022 года). [↑](#footnote-ref-1)
2. Платформа, содержащая более 5 000 000 книжных памятников в формате .pdf.

   НЭБ Книжные памятникию URL: https://kp.rusneb.ru (Дата обращения: 07 июня 2022 года). [↑](#footnote-ref-2)
3. Подробнее о специальных терминах см. раздел «Методология», с. 8. [↑](#footnote-ref-3)
4. Все изображения собраны в Приложении 2. [↑](#footnote-ref-4)
5. Public AI models in Transkribus. URL: <https://readcoop.eu/transkribus/public-models/> (Дата обращения: 7 июня 2022) [↑](#footnote-ref-5)
6. Этот раздел статьи в значительной степени основывается на книге Шицгала А.Г. «Русский типографский шрифт: вопросы истории и практика применения». Шицгал Абрам Григорьевич – историк и филолог, специализирующийся на истории русской типографики, один из крупнейших специалистов в этой области. [↑](#footnote-ref-6)
7. Сводный каталог русской книги гражданской печати XVIII века 1725-1800: В 5 т. Т. V. Указатели. М.: Книга, 1967. С. 278-290. [↑](#footnote-ref-7)
8. Шицгал А. Г. Русский типографский шрифт: вопросы истории и практика применения. М.: Книга, 1985. С. 61-71. [↑](#footnote-ref-8)
9. Muehlberger, Guenter, et al. "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study." Journal of documentation (2019) URL: <https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html> (Дата обращения: 12 июня 2022 года). [↑](#footnote-ref-9)
10. Milioni, N. *Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars*. Department of ALM. Theses within Digital Humanities, 2020. p. 13. [↑](#footnote-ref-10)
11. Milioni, N., p. 13. [↑](#footnote-ref-11)
12. Там же. [↑](#footnote-ref-12)
13. О том, как рассчитывается CER можно почитать на сайте платформы: URL: https://readcoop.eu/glossary/character-error-rate-cer/ (Дата обращения: 8 июня 2022 года). [↑](#footnote-ref-13)
14. Milioni, p. 14. [↑](#footnote-ref-14)
15. Milioni, p. 21. [↑](#footnote-ref-15)
16. Национальный корпус русского языка. URL: https://ruscorpora.ru/ (Дата обращения: 11 июня 2022 года). [↑](#footnote-ref-16)
17. Шицгал, с. 77. [↑](#footnote-ref-17)
18. [↑](#footnote-ref-18)