

Виктор А. БАРАНОВ

Кафедра лингвистики

Ижевский государственный технический университет  
имени М. Т. Калашникова, Ижевск, Россия

victor.a.baranov@gmail.com

UDK 81'324

003.349.1

Izvorni znanstveni članak

Primljen: 12. lipnja 2018.

Prihvaćen: 11. lipnja 2019.

**К ВОПРОСУ ОБ ИСПОЛЬЗОВАНИИ  
СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ ПОИСКА  
КОЛЛОКАЦИЙ И КОЛЛИГАЦИЙ В ДРЕВНЕЙШИХ  
СЛАВЯНСКИХ ТЕКСТАХ (НА МАТЕРИАЛЕ  
ГЛАГОЛИЧЕСКИХ РУКОПИСЕЙ КОРПУСА  
»МАНУСКРИПТ«)**

Работа посвящена вопросам методики поиска цельных сочетаний в коллекции глаголических текстов исторического корпуса *Манускрипт: славянское письменное наследие* (manuscripts.ru) и оценки их устойчивости. Демонстрируются возможности модуля n-грамм, позволяющие извлекать сочетания, имеющие разное количество компонентов, различную частотность, состоящие из текстовых форм или из лемм. Анализируются биграммы и триграммы, извлеченные с помощью статистической меры *Mutual Information* и встречающиеся одновременно в нескольких рукописях коллекции.

Особое внимание обращается на n-граммы, имеющие высокие статистические значения меры. В соответствии с особенностями меры наибольшие значения имеют редкие в коллекции сочетания. Анализ таких биграмм на основе словоформ позволил выявить устойчивые грамматические конструкции – коллигации. Показывается, что триграммы, состоящие из текстовых форм, являются не только грамматически, но и семантически цельными – коллокациями. Разнообразны биграммы с компонентами-леммами – предложно-падежные сочетания, именные сочетания с зависимыми притяжательными местоимениями и иные атрибутивные конструкции, глагольные конструкции с зависимыми существительными и другие. Анализ этих групп позволил выявить как коллигации, так и коллокации. Наиболее результативным стало извлечение триграмм на основе лемм – подавляющая часть первых нескольких десятков сочетаний, имеющих максимальное значение в соответствии со статистической мерой MI, представляет собой грамматически и семантически цельные конструкции или их части.

Делается вывод об эффективности использования статистических методов для извлечения коллокаций и коллигаций из корпусов, содержащих средневековые славянские рукописи. Указывается, что для комплексного решения поставленной задачи должны быть использованы различные виды n-грамм – двукомпонентные и трехкомпонентные, на основе текстовых форм и лемм, со свободным и фиксированным порядком следования компонентов. Наличие повторяющихся в различных рукописях грамматически и семантически цельных сочетаний позволяет сделать вывод о надтекстовой природе таких сочетаний.

Ключевые слова: текстовый корпус »Манускрипт«, manuscripts.ru, глаголическая рукопись, лингвистическая статистика, модуль n-грамм, коллокация, коллигация

## 1. ВВЕДЕНИЕ: КОРПУС ТЕКСТОВ. СТАТИСТИЧЕСКИЕ МЕТОДЫ. ГЛАГОЛИЧЕСКИЕ РУКОПИСИ

Корпусные технологии предоставили лингвистам новые возможности работы с текстами, главная из которых – наблюдение за функционированием лингвистических единиц (их активностью, сочетаемостью, вариативностью) в очень больших объемах текстов с учетом жанровых, временных, гендерных и иных характеристик произведений. Появление машиночитаемых размеченных коллекций средневековых письменных памятников, снабженных средствами поиска и упорядочения лингвистических данных, дает специалистам по истории славянских языков новый инструментарий как для нахождения материала для анализа традиционными методами, так и для постановки и решения задач, данными для которых являются количественные и дистрибутивные характеристики единиц текста.

Информация о сочетаемости является одной из важнейших характеристик текстовых форм и слов. В разной степени грамматически и семантически цельные сочетания являются объектом теоретических исследований морфологов, синтаксистов, фразеологов, лексикологов, семасиологов, а также материалом для создания систем автоматического перевода, распознавания речи, алгоритмов стилеметрии, процедур извлечения информации и разработок в области искусственного интеллекта.

Достаточно продуктивным способом нахождения устойчивых грамматических и/или семантических комплексов и сочетаний, претендующих на такой статус – коллокаций и коллигаций – считаются статистические методы оценки степени связанности компонентов текста.<sup>1</sup> Оценка

---

<sup>1</sup> О понятиях коллокация и коллигация и о некоторых статистических методах их выделения см., например: НОНЛОВА 2008; КОЅЕТКОВА 2018; ONAL 2017; PECINA;

осуществляется на основе данных о количестве некоторого сочетания, а также каждого из его компонентов в текстовом корпусе. Отклонение от средних, ожидаемых частот в сторону бóльших величин является основанием считать, что хотя бы один из компонентов имеет ограниченную сочетаемость, а сочетание в целом является несвободным. Подобный подход позволяет выявить, с одной стороны, наиболее частотные сочетания – синтаксические конструкции, клише, терминологические сочетания, характеризующие документ, коллекцию, подкорпус в целом, а с другой – редкие, единичные устойчивые конструкции с ограниченной сочетаемостью компонентов – имена собственные, термины, идиомы, характеризующие тему текстов (HOHLOVA 2008: 347–348).

Появления первых машиночитаемых коллекций средневековых славянских письменных памятников создали условия для экспериментов по применению корпусных методов к новому материалу (BARANOV 2015).

## 2. ИСТОРИЧЕСКИЙ КОРПУС И СТАТИСТИКА: ЦЕЛЬ, ЛИНГВИСТИЧЕСКИЕ ДАННЫЕ, МЕТОДЫ, ИНСТРУМЕНТАРИЙ

### 2.1. Цель работы

Работа является продолжением начатых автором опытов по выявлению в глаголических текстах XI века устойчивых сочетаний. В статье содержится изложение результатов экспериментов по поиску *повторяющихся в нескольких списках* коллокаций и коллигаций и установлению степени их устойчивости.

### 2.2. Лингвистические данные

Материалом для анализа послужила коллекция глаголических рукописей портала »Манускрипт: славянское письменное наследие« (manuscripts.ru), включающая транскрипции 3 кодексов и 5 фрагментов, а именно: 1. *Ассеманиево Евангелие, Slav. 3 Glag.*, X/XI вв. (ЕвАс); 2. *Псалтырь Димитрия Синайского, Cod. Sin. slav. 3/N*, третья четверть XI в. (ПсДС); 3. *Литургиарий, Cod. Sin. slav. 5/N*, вторая четверть XI в. (ЛитС);

---

SCHLESINGER 2006; VLAVACKAÂ 2011; ZHANG; IRIA; BREWSTER; CIRAVEGNA 2008.

4. *Литургиарий*, фрагмент, вторая четверть XI в. (ЛитСФр); 5. *Медицинские листки*, *Cod. Sin. slav. 3/N*, последняя четверть XI в. (МедЛ); 6. *Миня синайская*, *Cod. Sin. slav. 4/N*, последняя четверть XI в. (МинС); 7. *Киевские листки*, отрывок, X/XI в. и XII в. (КЛ) и 8. *Часослов*, фрагмент, первая четверть XI в. (Час) – (см. *Источники*).

Объем коллекции – 89172 формы.<sup>2</sup> Количество лемматизированных текстовых прецедентов – 55220, количество лемм – 3024.

Материал этой коллекции уже был объектом рассмотрения в статье (BARANOV 2017), в которой проанализированы данные 3-х кодексов и 4-х фрагментов.<sup>3</sup> В работе показаны возможности модуля n-грамм корпуса «Манускрипт» (URL: [http://manuscripts.ru/mns/cred\\_ngr.stat](http://manuscripts.ru/mns/cred_ngr.stat)) и описано несколько экспериментов, которые позволили выявить многокомпонентные повторяющиеся чтения и формулы и двухкомпонентные семантически и грамматически устойчивые сочетания. Последние найдены с помощью статистических мер *Mutual information*, *T-Score* и *Dice*. В частности, сделан вывод, что набор коллокаций индивидуален для каждого кодекса. И только в нескольких случаях отмечено использование пятикомпонентных и четырехкомпонентных сочетаний одновременно в нескольких рукописях.<sup>4</sup>

Анализ трехкомпонентных и двухкомпонентных устойчивых конструкций, использующихся в нескольких рукописях коллекции, не был осуществлен. Но именно такие сочетания, характеризующие не один, а несколько разножанровых текстов, могут рассматриваться как существенная характеристика подкорпуса в целом.

<sup>2</sup> Объем рукописей и отрывков: ЕвАс – 43 287; ПсДС – 30 659; ЛитС – 12 474; КЛ – 1228; МинС – 485; МедЛ – 279; Час – 218; ЛитСФр – 542.

<sup>3</sup> В коллекции отсутствовал фрагмент *Литургиария* (Санкт-Петербург, БРАН 24.4.8 & РНБ, *Глаг. 2*), втор. четв. XI в., 3 л.

<sup>4</sup> Одна пятиграмма – на нѣво и на зѣмля 2х: ПсДС: на ѿвѣсѣ і на зѣмѣ 116.1.1–20, ЕвАс: на нѣво и на зѣмѣ 73.2.2–3 (все примеры даются в упрощенной записи); 29 четырехграмм (в ряде случаев – части более развернутых конструкций) с разной структурой и разной степенью грамматической и семантической цельности: парные формулы – [вѣ илиа] оца и сѣна и, нхѣ денѣ и ноцѣ, нѣнѣ и присно ѿ, отѣ вѣстокѣ і западѣ, подчинительные конструкции, в том числе с предложно-падежными компонентами, – вѣ домѣ оца моего, вѣнѣ іѣ бѣ іѣвѣ, вѣ дѣни вѣ наже, во вѣса прѣдѣлы іхѣ, і вѣсѣ ѣже сѣтѣ, глагольно-предложные и беспредложные конструкции: Вѣзнесѣ сѣ на невеса, ѿ вѣзврати сѣ вѣ, еси на главѣ еѣго, ѿ весела сѣ вѣ, прѣданѣ вѣдетѣ вѣ рѣцѣ, і вѣзвратѣтѣ сѣ вѣспѣтѣ, і поклонѣтѣ сѣ еѣнѣ и некоторые другие (далее = нек. др.).

### 2.3. Инструменты и методика

Выборки для анализа были подготовлены с помощью модуля n-грамм корпуса (BARANOV 2014; BARANOV 2016.a; BARANOV 2016.b). Параметры n-грамм задавались с помощью полей: »Компонент« – количество компонентов n-граммы; »Тип компонентов« – тип лингвистической единицы (словоформа или лемма); »Границы конструкций« – учет или не учет разделителей контекста; »Следование« – фиксированный или свободный порядок компонентов; »Мера« – статистический метод оценки сочетания.

В работе для нахождения характерных для коллекции устойчивых сочетаний использована статистическая мера MI – *Mutual information* («взаимная информация»).<sup>5</sup> Как известно, она присваивает наибольшие значения уникальным сочетаниям. Недостатком меры MI является завышение значения редких сочетаний, содержащих единично встречающиеся компоненты, что корректируется, в частности, установлением порога частотности n-грамм (ÂGUNOVA; PIVOVAROVA 2010; ÂGUNOVA; PIVOVAROVA 2013; LUKAŠEVIČ; LOGAČEV 2010; ZALESSKAÂ 2014: 284; ZAHAROV; HOHLOVA 2014: 291; KOŠETKOVA).

## 3. ОБСУЖДЕНИЕ: N-ГРАММЫ. КОЛЛОКАЦИИ. КОЛЛИГАЦИИ

### 3.1. Биграммы. Словоформы

Параметры n-грамм: тип компонентов – словоформы, количество компонентов – 2, количество текстовых прецедентов –  $\geq 2$ , следование – фиксированное, границы конструкций – учтены.

Результат – 6553 биграммы (69839 текстовых прецедентов). Наибольшее значение меры – 15,444 (врдж (2) арүнж (2) ПсДС 2×), наименьшее – -5,104 (2 и (4551) ѵ (1348) ЛитС 1×, 1 ЕвАс 1×). Проанализированы начальные 250 сочетаний, зафиксированные в двух и более текстах и имеющие значения в соответствии с мерой MI от 13,859 до 5,140. Пере-

<sup>5</sup>  $MI = \log_2 \frac{F(w_1, w_2) \times N}{F(w_1) \times F(w_2)}$ , где  $w_1$  – первый компонент,  $w_2$  – второй компонент,  $F(w_1)$ ,  $F(w_2)$  – абсолютные (независимые) частоты  $w_1$  и  $w_2$ ,  $F(w_1, w_2)$  – абсолютная частота n-граммы в подкорпусе,  $N$  – общее число словоформ в корпусе; см., например: EVERT 2004.

чень групп биграмм, расположенных по убыванию количества примеров в группе, приведен в *Приложении 2*.

Выборка и систематизация биграмм, использующихся в различных текстах коллекции, позволяет сделать несколько наблюдений.

Некоторые из биграмм представляют собой сочетания, компоненты которых используются друг с другом чаще, чем с другими компонентами (значения биграмм достаточно высоки – до 13 статистических единиц): *бес порока*,<sup>6</sup> *очима моима*, *грѣхъ нашихъ*, *ты еси*, *г(оспод)а б(ог)а*, *другъ друга*, *всему миру*, *живи будете*, *раби недостоини*, *дньесьнии день*, *съжежетъ огнемъ*, *простъръ руку* и некоторые другие, именно они являются устойчивыми или могут рассматриваться в качестве претендентов на устойчивые.

Некоторые из групп – *предложно-надежные сочетания* с существительными или местоимениями, *существительное + притяжательное местоимение* – очень многочисленны и с точки зрения лексического состава биграмм, и с точки зрения их использования в корпусе.

Между моделью сочетания, количеством биграмм перечня, построенных по этой модели, и степенью связанности компонентов нет зависимости: устойчивыми являются как биграммы, единичные в своей группе – *другъ друга*, *съжежетъ огнемъ*, *раби недостоини*, *дньесьнии день*, *живи будете* и др., так и входящие в группу с большим количеством примеров – *бес порока*, *из устъ*, *подъ криль*, *очима моима*, *люди своя*.

Между тем биграммы некоторых групп тяготеют к большим или меньшим значениям меры MI; например, имеют:

1. высокие значения две биграммы по модели *глагол + существительное* (*съжежетъ огнемъ*, *простъръ руку*), все шесть биграмм с компонентом *весь* (*весь миръ*, *въсеи земли* и др.), три сочетания существительного с зависимым прилагательным (*раби недостоини*, *дньесьнии день*, *вечьнии б(ож)е*), предикативное *сигѣтъ свѣтъ*, личное *Господь Богъ*, а также конструкции из двух существительных (*свѣтъ лица*, *другъ друга*);

2. низкие – биграммы по моделям *личное местоимение + быти* (*ты еси*, *азъ есмь* и др.), *глагол + зависимое местоимение* (*помози намъ*, *въздасть тебѣ* и др.), глагольно-предложные сочетания (*въниду въ*, *възиде на* и др.), перфектные формы глагола (*далъ еси*, *възвеличилъ еси*), конструкции с начальными *да* (*да омочить*, *да потрѣблю* и др.), *яко* (*гько капль*, *гько мръत्वъ*, *гько благъ*) или частицами *не* (*не могутъ*, *не познаши*

<sup>6</sup> В основном тексте работы примеры даны в упрощенном написании.

и др.) или же (*онъ же, тѣмъ же* и др.), конструкции с послелогом *ради* (*чесо ради, сего ради*), сочетания глагол + *ся* (*сѣбудеть ся, яви ся* и многие другие – далее = мн. др.);

3. преимущественно низкие – предложно-падежные сочетания с существительными (*въ глубину, въ гробѣхъ, въ храмѣ, въ вѣкѣ, на одрѣ, на горахъ, отъ страха, отъ лица, прѣжде вѣкѣ* и мн. др.) и местоимениями (*по сихъ, на нь, къ тебѣ* и мн. др.), за единичными исключениями – *до коньца, прѣдъ очима, подѣ криль, бес порока, прѣдъ тобою, ис тебе* и некоторые др.;

4. существенно различающиеся в группе значения – сочетания *существительное + притяжательное местоимение*: 4.1. низкие – с *его* (*главь его, матери его, нозѣ его* – от 5,3 до 5,9); 4.2. в основном средние или высокие – с местоимениями *твои* (*милость твою, десница твоѣ, раба твоего, вѣра твоѣ* и мн. др. – от 6,2 до 10,1), *свои* (*руцѣ свои, лице свое* и др. – от 6,5 до 8,9), *мои* (*тело мое, очи мои, ср(ѣ)д(ѣ)це мое* и др. – от 6,0 до 9,2; *очима моима* – 13,9), *нашь* (*грѣхъ нашихъ, б(ог)ѣ нашъ* и др. – от 5,8 до 9,7).

Основное количество биграмм составляют структуры, один из компонентов которых – служебное слово или местоимение, и только очень малое количество сочетаний двух знаменательных слов, репрезентирующих сложное понятие.

Сказанное позволяет сделать выводы о том, что выделенные группы биграмм, в первую очередь – с большим количеством примеров, представляют собой грамматические конструкции, построенные по некоторым моделям, заполнение позиций которых словами достаточно свободно, а сами биграмм являются реализацией некоторых достаточно устойчивых и постоянно воспроизводимых в списках глаголической коллекции конструкций, обладающих тяготеющими друг к другу в текстах грамматическими характеристиками – коллигациями.<sup>7</sup> Одновременно с этим некоторые из биграмм на основе словоформ, компоненты которых сочетаются чаще, чем с другими словами (что проявляется в высоких значениях статистической меры MI), должны быть отнесены к статистически (коллокации), а в ряде случаев – и лингвистически (речевые формулы, обороты) устойчивым.

<sup>7</sup> »Коллигация – это комбинация лексических и грамматических характеристик, имеющих тенденцию к совместной встречаемости« (КОЏЕТКОВА: 302).

### 3.2. Триграммы. Словоформы

Параметры n-грамм: тип компонентов – словоформы, количество компонентов – 3, количество текстовых прецедентов –  $\geq 2$ , следование – фиксированное, границы конструкций – учтены.

Результат – 2226 триграмм (52 234 текстовых прецедентов). Наибольшее значение меры – 15,444 (*зълълхъ* (3) *окоусънхъ* (2) *мъсл[.]* (2) ЛитС 2×),<sup>8</sup> наименьшее – -17,070 (*же* (1185) и (4551) и (4551) ЕвАс 2×). Проанализированы все сочетания, встречающиеся одновременно в двух или более текстах; их значения в соответствии с мерой MI от 6,567 до -15,396.

Всего триграмм, встречающихся в нескольких источниках, – 47. Все они приведены в таблице 1 *Приложения*.

Выборка позволяет сделать следующие заключения.

Доля триграмм, одновременно использующихся в различных текстах коллекции, по сравнению с общим количеством триграмм в выборке крайне мала. Триграммы, имеющие достаточно высокое статистическое значение, представляют собой семантически и грамматически цельные сочетания (*сигеть свѣтъ лица, г(оспод)а б(ог)а нашего* и т. д.). Основными структурами первой половины списка являются именные (*въ дъ-несънии день, по въсеи земли* и др.) и глагольные (*въниду въ домъ, будетъ въ рущь* и др.) с зависимыми компонентами различной грамматической семантики, а также одна предикативная конструкция (*онъ же рече*). Структуры второй части перечня представляют собой сочетания грамматически связанных (*да не будетъ*) и/или контекстно обусловленных (*и присно и*) компонентов. Триграммы, располагающиеся в самом конце списка выборки, – сочетания находящихся в контакте служебных слов с очень большой частотностью (*ся* (1975) и (4551) *не* (1283)). Большая часть повторяющихся в разных текстах триграмм встретила только по два раза – в больших по объему ЛитС и ЕвАс или ПсДС и ЕвАс. Различна семантическая цельность компонентов сочетаний: от имени *Господь Богъ*, регулярно встречающегося в *Псалтыри – свѣтъ лица* (Пс. 4:7, 43:4, 88:16 и др.), конструкций с предопределенным грамматическим компонентом (*въниду в домъ, и възиде на*), конструкций с конкретизирующими компонентами (*въ весь миръ, въ вргмя свое*) и формул (*онъ же*

<sup>8</sup> Где (2) – количество форм в подкорпусе, 2× – количество употреблений n-граммы в подкорпусе.

рече) – до свободных (*будетъ въ рече*) и контекстно обусловленных (*а не въ*) сочетаний. Провести границу между устойчивыми и свободными сочетаниями в перечне невозможно, расчлененность конструкций постепенно возрастает от начала к концу выборки.

Несмотря на небольшое количество семантически и грамматически цельных триграмм, их наличие позволяет говорить об определенной степени близости текстов глаголической коллекции друг другу на уровне коллокаций.

### 3.3. Биграммы. Леммы

Параметры n-грамм: тип компонентов – леммы, количество компонентов – 2, количество текстовых прецедентов –  $\geq 2$ , следование – свободное, границы конструкций – учтены.

Результат – 3671 биграмма (29508 текстовых прецедентов). Наибольшее значение меры – 14,859 (*вдѣланиѣ* (3) *врачѣнѣ* (2) ЕвАс 2×), наименьшее – -6,956 (*въ* (4705) *въ* (4705) ЕвАс 2×). Проанализированы начальные 250 сочетаний, зафиксированные в двух и более текстах; имеют значения в соответствии с мерой MI от 10,900 до 3,458.

Основными группами перечня являются:

1) предложно-падежные сочетания с существительными: *за умножение*,<sup>9</sup> *до коньць* (*до коньца*),<sup>10</sup> *до западъ*, *из уста* (*из усть*), *подъ нога* (*подъ ногама* и др.), *до скотъ* (*до скота*), *на погребение*, *отъ юность* (*отъ юности*), *отъ въстокъ*, *на въстание*, *на лоно* (*на лонѣ*, *на лоно*), *на небо*, *отъ западъ*, *отъ лице* (*отъ лица*), *отъ печаль* (*отъ печали*), *къ гробъ* (*къ гробу* и др.), *на вечеръ*, *отъ земля*, *отъ страхъ* (*отъ страха*), *до небо* (*до небеси* и др.), *на крило* (*на крилѣ* и др.), *на одръ*, *на кровь* (*на кровѣхъ* и др.), *на мѣсто*, *отъ гнѣвъ* (*отъ гнева*), *на ложе*, *на прѣстолъ* (*на прѣстолѣ*), *въ прѣполовление*, *въ студеньць*, *отъ адъ* (*отъ ада*), *отъ стѣть* (*отъ стѣти*), *отъ туждъ* (*отъ туждихъ*), *отъ чръво* (*отъ чръва*), *по глава* (*по главѣ*), *въ оставление*, *въ раи*, *на камы* (*на камени*), *прѣдъ лице*, *съ радость* (*съ радостию*), *къ цръкы* (*къ цр(ь)к(ь)ви*), *прѣдъ око* (*прѣдъ очима*), *въ притьча* (*въ притьчахъ*), *отъ старьць* (*отъ старьць* и

<sup>9</sup> Сочетание лемм приводятся в порядке уменьшения значения меры MI.

<sup>10</sup> В скобках для наглядности приводятся одна-две текстовые формы; при этом отсутствие после примера(ов) и в конце перечней «... и др.» означает, что приведены все грамматические конструкции конкретной биграммы и все биграммы данной группы.

др.), *въ дворъ, въ сласть, отъ градъ (отъ града), на поле (на поли и др.)* – и местоимениями: *къ ть (къ тому), по ть (по томъ), о съ (о семь), у ты (у тебе), къ ты (къ тебѣ)*;

2) именные сочетания с зависимыми притяжательными местоимениями: *немоць нашъ (немоць нашу и др.), грѣхъ нашъ, съкровище свои (съкровища своего и др.), отьць нашъ (отьци наши и др.), юность мои (юности моя и др.), заповѣдь твои (заповѣди твоя и др.), ольтарь твои, прѣстъ свои, рабъ твои, люди свои, благость твои (благости твоя и др.), душа мои (души моя и др.), пажить твои (пажити твоя и др.), глаголъ мои, братъ твои, милость твои (милость твою и др.), истина твои (истинѣ твоей и др.), плоть мои (плоти мои и др.), срьдьце вашъ (срьдцье вашихъ), нога мои (ногама моима и др.), мои слъза (моима слъзы), мръжа свои (мръжу свою и др.), слово твои (слова твоя и др.), домъ свои, дѣло свои (дѣла своихъ), око мои (оци моима и др.), богъ нашъ, гнѣвъ твои, прѣстъ мои, срьдьце свои (срьдцье своего и др.), душа вашъ (души ваша), имя твои (имени твоего и др.), лице твои (лица твоего и др.), рабъ свои, нога нашъ (ноги наши и др.), душа свои (души свою и др.), чръво твои (чръва твоего), гласъ мои, отьць мои, поношение мои (поношение мое), хръстость твои, братъ мои, душа нашъ (души наша и др.), уста твои, село свои (село свое и др.), страхъ твои, свои градъ, гнѣвъ свои*;

3) сочетания существительного с зависимым существительным: *подъножие нога (подъножие ногама и др.), крѣвь хръстость (крѣвь хръста и др.), другъ другъ (другъ друга), домъ симонъ (дому симона), свѣтъ лице (свѣтъ лица), мать сынъ (сынь матери, мати сыну), домъ отьць (дому отьца), днь животъ (дни живота), душа рабъ (души раба и др.)*;

4) сочетания существительного с зависимым прилагательным: *симонъ прокаженъ (симона прокаженаго), рабъ недостоинъ, животъ вѣчьнъ, истинънъ богъ (истиннааго бога), богъ живъ (бога жива, бога живааго)*;

5) глагольные конструкции с зависимыми существительными: *насадити виноградъ, насадити десница, сътворити мьсть, възвести око (възвѣсь очи), сътворити плодъ, гсти хлѣбъ, творити добро, прославити имя, съпасти рабъ (спаси раба и др.), посылати духъ (пошли души и др.), видѣти дѣло (видѣши дѣла и др.), сътворити добро (добро сътвори и др.)*, а также сочетание отглагольного существительного с зависимым существительным: *оставление грѣхъ (оставление грѣховъ, оставление грѣхомъ)*;

6) глагольные конструкции с зависимыми личными местоимениями: *приносити ты (приносимъ тебѣ и др.), привлеци азъ (привлѣци мене и др.), угодити ты (угождѣщиими тебѣ), служити азъ (служить мнѣ и др.), одрѣжати азъ (одрѣжи мене и др.), поклонити ты (поклонить ти и др.), утвердити мы (утверди насъ и др.), въсхвалити ты (въсхвалятъ ты и др.), помилувати азъ (помилуй мя), уготовати вы (уготованое вамъ), достояти ты (достойтъ тебѣ), мстити азъ (мести мнѣ и др.), въздати ты (въздамъ тебѣ и др.), слушати азъ (слушаи мене и др.), съпасти азъ (с(ъ)п(а)си мя), даяти ты (тебе даи, тебѣ дающе и др.);*

7) глагольно-предложные сочетания: *въстати на (въстающимъ на и др.), приити къ, възложити на, призрьгати на, съдѣгати на, навести на, съдѣгати о, прѣгати на, ходити по, възвести на, съходити на, посылати къ;*

8) аналитические грамматические формы: *быти длѣжьнѣ (бѣ длѣжьнѣ и др.), прискрѣбнѣ быти (прискрѣбнѣ еси и др.), умолити быти (умоленѣ буди), подобнѣ быти (подобно естѣ), раздрушити быти (раздрушилъ еси и др.), достоинѣ быти, проданѣ быти (проданѣ бысть и др.), възвеличити быти (възвеличилъ естѣ), възложити быти (въложилъ еси и др.), вѣренѣ быти (вѣренѣ естѣ и др.);*

9) конструкции с местоимением *въсь*: *въсь мирѣ, въсь елико, въсь прѣдѣль, въсь съ (въсьми сими и др.), въсь лѣто (въсего лѣта и др.), въсь стадо (всему стаду и др.), въсь люди (вся люди и др.), въсь колѣно (въсь колѣна);*

10) глаголы с *ся*: *възиграти ся, просльзити ся, състарити ся, убояти ся, плакати ся, поклонити ся, боити ся, възвратити ся, отвѣрещи ся, оправѣдити ся, исповѣдати ся, наплѣнити ся, прѣклонити ся, облеци ся, уклонити ся, кляти ся, съкончати ся, възнести ся, въселити ся, исплѣнити ся, възносити ся, отѣяти ся, утѣшити ся, обратити ся, прославити ся.*

Иные группы малочисленны: глаголы в повелительном наклонении с существительным в звательной форме – *помянути господь (помяни г(оспод)и), помилувати господь (помилуй г(оспод)и);* предикативные конструкции – *рещи господь (рече г(оспод)ь), азъ въскрѣсити (азъ въскрѣшу)* и некоторые другие.

Достаточно большое количество биграмм представляет собой несвязанные грамматически или семантически конструкции: *ради помазати (ради помаза), ся горько, свои настави* и др.

Особый интерес представляют подчинительные конструкции, в которых оба компонента выражены знаменательными частями речи и пред-

ставляют собой в достаточно высокой степени цельные именованя лиц и речевые формулы (группы 3, 4), а в группе 5 – конструкции с наиболее актуальной для текстов лексико-семантической связью глаголов и зависимых существительных.

Устойчивость этих примеров доказывается не только смысловой и грамматической связью компонентов и использованием их в нескольких текстах коллекции, но и ограниченной сочетаемостью многих из них в рукописях:

*симонъ прокаженъ / курингънинъ / ионинъ / нарицаемъ / етеръ / свои;  
животъ въчьнъ / бесконъчьнъ / мои / твои / нашъ;  
рабъ недостоинъ / тышитъ / лукавъ / зълъ / гръшьнъ / проданъ / бо-  
ляшитъ / добръ / блаженъ / благъ / върънъ / господьнъ / свои / твои  
/ мои / инъ / ть.*

Глагол *сътворити* и его формы в коллекции имеют, кроме присутствующих в выборке *мьсть*, *плодъ*, *добро*, относительно небольшой круг и других зависимых слов – *свѣтило*, *брание*, *погребение*, *избавление*, *бракъ*, *память*, *телець*, *чудо*, *льсть*, *любы*, *небо*, *часъ*, *зъло*, *врѣмя*, *животъ*, *благо*, *свѣтъ*, *чловѣкъ*, *милость*, *ничъсоже*, *всья*. При этом сочетания с *мьсть*, *плодъ*, *добро* с глаголом *сътворити* имеют наибольший статистический вес по сравнению с другими. Глагол *насадити* имеет только три зависимых существительных – *корень*, *виноградъ* и *десница*, из них два последних – среди слов, с которым *насадити* используется в нескольких рукописях коллекции. Глагол *възвести* – два зависимых существительных – *очи*, *хлѣбъ*; *гъсти* два – *мана*, *хлѣбъ*; *съпасти* три – *душа*, *рабъ*, *люди*; *прославити* три – *имя*, *сынъ*, *отьць*; *творити* несколько больше – *делание*, *шествие*, *безаконие*, *плодъ*, *чюдеса*, *знамение*, *добро*, *дѣло*, *слово*, *дѣло*, *милость*; *посълати* также – *мости*, *сытость*, *вои*, *избавление*, *источникъ*, *стрѣла*, *брашьно*, *духъ*, *свѣтъ*, *сила*, *слово*, *рабъ*, *сынъ*, *исусъ*; *видѣти* также – *видѣние*, *юноша*, *камень*, *огнь*, *съмръть*, *исусъ*, *хръстость*, *безаконие*, *коньць*, *лице*, *гробъ*, *море*, *духъ*, *свѣтъ*, *чловѣкъ*, *отьць*, *ученикъ*, *сынъ*.

Высокие статистические значения сочетаний других групп свидетельствует об активности некоторых грамматических моделей – коллигаций »предлог + существительное«, »глагол + предлог« и др., а лексический состав – о повторяемости их лексической реализации в разных глаголических рукописях.

### 3.4. Триграммы. Леммы

Параметры n-грамм: тип компонентов – леммы, количество компонентов – 3, количество текстовых прецедентов –  $\geq 2$ , следование – свободное, границы конструкций – учтены.

Результат – 1973 триграммы (15547 текстовых прецедентов). Наибольшее значение меры – 9,159 (триговати (13) съдравиѣ (4) врачь (6) ЕвАс 2×), наименьшее – -22,491 (и (15437) и (15437) не (15437) ПсДС 6×). Проанализированы начальные 100 сочетаний, зафиксированные в двух и более текстах; имеют значения в соответствии с мерой MI от -0,686 до -10,770. Несмотря на низкие статистические значения этих триграмм (а соответственно расположение в полной выборке в 7-м десятке), находящиеся в позициях 1–50 n-граммы, встречающиеся в нескольких текстах, представляют собой семантически и грамматически цельные конструкции или их части:

предикативные – *иже насади десница, онъ же рече, нгсмь бо достоинь, съказа мьнь г(оспод)ь, въси напишуть ся;*

именные – *плодь чръва твоего, слово твоея истины / словесь твоихъ истина, въ весь мирь, вся дьни живота, душа раба твоего, г(осподи) милость твою, [глаголи] мирь на люди, г(осподи) б(ог)ъ нашъ, пргьдъ въсьми людьми, подъ нозь его, въ ногахъ нашихъ, [помѣни] сънемъ твои иже [сътяжа];*

глагольные – *очима моима слъзы [точити], възвратятъ ся въспятъ, утвьрди силою твоею, [ангеломъ] своимъ заповѣсть о тебѣ, еси къ тому, пришедъ въсели ся, исповѣдаю ся тебѣ, поклони ти ся, еси на н(е) б(е)сехъ, помилуй мя г(осподи), придетъ къ тебѣ, нгьсть бо д(оу)ха, уготоваль еси пргьдъ [мьною], въложилъ еси въ;*

именные парные – *слъньце и луна, въстокъ и западъ, [въ] родъ и родъ, гора и хлъмъ, дьнь и ноцъ, петра и павла, нынъ и присно;*

наречные – *и коль красно;*

предложно-падежные – *и за умножение [безаконя], по срьдъ же, и на въстание;*

иные – *[ц(ѣ)са)рь] велеи по всѣи [земи], къ тому и, ся къ цр(ь)к(ъ)ви, ни единъ отъ.*

Просмотр позиций 51–100 выборки добавил к сочетаниям немного: *въ дому о(ть)ца, въ срьдци своемъ, сътворилъ еси намъ, и плодь сътвори, въ всемъ мирь, о(ть)ца и с(ы)на, г(оспод)и б(о)же мои.*

Использование параметра »Следование« запросной формы с целью нахождения сочетаний, компоненты которых могут находиться в любой последовательности друг к другу, привело в ряде случаев к объединению к од-

ной триграмме разных моделей: *въ прѣстѣ свою / прѣста своего въ, съказа мнѣ г(оспод)ь / съкажи мнѣ г(оспод)и, въ ногахъ нашихъ / ноги наша въ.*

Таким образом, выборка триграмм на основе лемм дает существенно иные результаты, чем выборка на основе текстовых прецедентов: подавляющая часть первых нескольких десятков сочетаний представляет собой грамматически и семантически цельные конструкции или их части. Использование этих конструкций в различных глаголических рукописях позволяет с большой долей вероятности считать их не свободными, а цельными, какие бы причины ни были тому причиной – интертекстуальные, речевые или языковые.

#### 4. ЗАКЛЮЧЕНИЕ

Извлечение биграмм и триграмм на основе текстовых форм и лемм из глаголической коллекции исторического корпуса «Манускрипт: славянское письменное наследие» ([manuscripts.ru](http://manuscripts.ru)) с помощью статистической меры *Mutual information* и их анализ дает право говорить о высокой информативности выборок для поиска с высокой степенью вероятности цельных сочетаний. Ограничение перечней анализируемых n-грамм примерами, встречающимися в нескольких рукописях, позволило не только выявить круг сочетаний, характеризующих коллекцию в целом, но и с большой степенью вероятности предположить, что цельность сочетаний, имеющих высокое статистическое значение, одновременно является и результатом их свойств в одном тексте, и, в связи с воспроизводимостью в нескольких, должна рассматриваться как общетекстовая и надтекстовая для проанализированных рукописей.

Анализ материала показал, что провести границу между цельными и свободными сочетаниями только на основании статистического значения невозможно. Тем не менее сочетания, имеющие максимальные значения в ограниченной определенными параметрами выборке (удаление единичных сочетаний, приведение словоформ к леммам) и повторяющиеся в нескольких текстах, следует с высокой долей вероятности признавать значимыми, цельность которых определяется семантической и грамматической связностью компонентов.

Варьирование параметров n-грамм позволяет обнаружить как сочетания с лексико-грамматической связностью компонентов – коллокации, так и морфолого-синтаксические модели, представленные в коллекции частотными реализациями, – коллигации.

Несмотря на то, что использование статистической меры *Mutual Information* для извлечения редких цельных сочетаний в коллекции глаголических рукописей показало высокую эффективность и результативность, окончательная оценка степени устойчивости большей части сочетаний или грамматических моделей может быть осуществлена с привлечением и других, собственно лингвистических методов.

## БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Лингвостатистический анализ однокомпонентных и многокомпонентных лексических единиц исторического корпуса “Манускрипт”» (проект № 18/012/00463)

## СОКРАЩЕНИЯ / ABBREVIATIONS

- БРАН / BRAN = Библиотека Российской академии наук, С.-Петербург. / Library of the Russian Academy of Sciences, St. Petersburg.  
РАН / RAN = Российская академия наук. / Russian Academy of Sciences.  
ААН / AAN = Австрийская Академия наук / Austrian Academy of Sciences.  
РНБ / RNB = Российская национальная библиотека, С.-Петербург. / National Library of Russia, St. Petersburg.  
НБУ / NBU = Национальная библиотека Украины имени В. И. Вернадского, Киев / National Library of Ukraine named after V. I. Vernadsky, Kiev

## ИСТОЧНИКИ

- ЕвАс = *Ассеманиево Евангелие*, Ватиканская (Апостолическая) библиотека, *Slav. 3 Glag.*, X/XI вв., 158 л. Подг. Р. Ф. ТУБИЛОВА; В. А. БАРАНОВ (Ижевский государственный технический университет). URL: [http://manuscripts.ru/mns/main?p\\_text=56673793](http://manuscripts.ru/mns/main?p_text=56673793) (15.5.2018).  
ПсДС = *Псалтырь Димитрия Синайского*, Синай/Египет, Монастырь св. Екатерины, *Cod. Sin. slav. 3/N*, третья четверть XI в., 145 л. Подг. А. Н. СОБОЛЕВ (РАН/ААН/Марбургский университет); М. МАКРОБЕРТ (Оксфордский университет); М. ГАУ (Венский технический университет); под ред. Х. МИКЛАСА (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=42> (15.5.2018).

- ЛитС = *Литургиарий* (ранее «Миссал»), Синай/Египет, Монастырь св. Екатерины, *Cod. Sin. slav. 5/N*, вторая четверть XI в., около 80 л. Подг. В. САДОВСКИ; Д. ХЮРНЕР; Ф. ВАНДЛ; Х. МИКЛАС, под ред. Х. МИКЛАСА (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=67> (15.5.2018).
- ЛитСФр = *Литургиарий*, фрагмент («Синайский служебник», «Листки Крылова и Успенского»), Санкт-Петербург, БРАН 24.4.8 (Срезн. 55) & РНБ, *Глаг. 2*, вторая четверть XI в., 3 л. Подг. Х. МИКЛАС (Венский университет); В. А. БАРАНОВ (Ижевский государственный технический университет); под ред. Х. МИКЛАСА (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=69> (15.5.2018).
- МедЛ = *Медицинские листки*, Синай/Египет, Монастырь св. Екатерины, *Cod. Sin. slav. 3/N*, посл. четв. XI в., 2 л. Подг. Х. МИКЛАС; Д. ХЮРНЕР, под редакцией Х. Микласа (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=59> (15.5.2018).
- МинС = *Минея синайская*, Синай/Египет, Монастырь св. Екатерины, *Cod. Sin. slav. 4/N*, последняя четверть XI в., 2 л. Подг. Х. МИКЛАС; Д. ХЮРНЕР (Венский университет); М. ГАУ (Венский технический университет); под ред. Х. Микласа (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=48> (15.5.2018).
- КЛ = *Киевские листки* («Киевский миссал»), НБУ, *ДА/П. 328*, X/XI в. и XII в., отрывок, 7 л. Подг. В. А. БАРАНОВ (Ижевский государственный технический университет). URL: [http://manuscripts.ru/mns/main?p\\_text=61350133](http://manuscripts.ru/mns/main?p_text=61350133) (15.5.2018).
- Час = *Часослов*, фрагмент, Синай/Египет, Монастырь св. Екатерины, первая четверть XI в., 1 л. Подг. Н. ГЛИБЕТИЧ (Еврейский университет, Иерусалим); под ред. Х. Микласа (Венский университет). URL: <http://manuscripts.ru/mns/portal.main?p1=66> (15.5.2018).

## ЛИТЕРАТУРА

- ÂGUNOVA; PIVOVAROVA 2010. = ЯГУНОВА, Е. В.; Л. М. ПИВОВАРОВА. 2010. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов. *Сборник научно-технической информации. Сер. 2, № 6*. URL: [http://medialing.spbu.ru/upload/files/file\\_1394529742\\_4311.pdf](http://medialing.spbu.ru/upload/files/file_1394529742_4311.pdf), [http://webground.su/services.php?param=priroda\\_collac&part=priroda\\_collac.htm](http://webground.su/services.php?param=priroda_collac&part=priroda_collac.htm) (15.5.2018). [ÂGUNOVA, E. V.; L. M. PIVOVAROVA. 2010. Priroda kollokacij v russkom âzyke. Opyt avtomatičeskogo izvlečeniâ i klassifikacii na materiale novostnyh tekstov. *Sbornik nauchno-tekhničeskoj informacii. Ser. 2, № 6*. URL: [http://medialing.spbu.ru/upload/files/file\\_1394529742\\_4311.pdf](http://medialing.spbu.ru/upload/files/file_1394529742_4311.pdf), [http://webground.su/services.php?param=priroda\\_collac&part=priroda\\_collac.htm](http://webground.su/services.php?param=priroda_collac&part=priroda_collac.htm) (15.5.2018).]

- ÂGUNOVA; PIVOVAROVA 2013. = ЯГУНОВА, Е. В.; Л. М. ПИВОВАРОВА. 2013. От коллокаций к конструкциям. *Русский язык: конструкционные и лексико-семантические подходы*. С. С. Сай (ред.). Санкт-Петербург: Труды Института лингвистических исследований РАН. URL: [https://www.academia.edu/4461901/ОТ\\_КОЛЛОКАЦИЙ\\_К\\_КОНСТРУКЦИЯМ](https://www.academia.edu/4461901/ОТ_КОЛЛОКАЦИЙ_К_КОНСТРУКЦИЯМ) (15.5.2018). [ÂGUNOVA, E. V.; L. M. PIVOVAROVA. 2013. Ot kollokacij k konstrukciám. *Russkij ázyk: konstrukcionnye i leksiko-semantičeskie podhody*. S. S. Saj (red.). Sankt-Peterburg: Trudy Instituta lingvističeskikh issledovanij RAN. URL: [https://www.academia.edu/4461901/ОТ\\_КОЛЛОКАЦИЙ\\_К\\_КОНСТРУКЦИЯМ](https://www.academia.edu/4461901/ОТ_КОЛЛОКАЦИЙ_К_КОНСТРУКЦИЯМ) (15.5.2018).]
- BARANOV 2014. = БАРАНОВ, В. А. 2014. Организация поиска и демонстрации коллокаций в корпусе «Манускрипт». *Проблемы истории, филологии, культуры* 3 (45): 275–277. [BARANOV, V. A. 2014. Organizaciâ poiska i demonstracii kollokacij v korpuse »Manuskript«. *Problemy istorii, filologii, kul'tury* 3 (45): 275–277.]
- BARANOV 2015. = БАРАНОВ, В. А. 2015. Исторический корпус как цель и инструмент корпусной палеославистики. *Scripta & e-Scripta 14–15*: 39–62. URL: <https://drive.google.com/file/d/0BwBejXXryRcROVQ4TnlpZFh6am8/view?usp=sharing> (15.5.2018). [BARANOV, V. A. 2015. Istoričeskij korpus kak cel' i instrument korpusnoj paleoslavistiki. *Scripta & e-Scripta 14–15*: 39–62. URL: <https://drive.google.com/file/d/0BwBejXXryRcROVQ4TnlpZFh6am8/view?usp=sharing> (15.5.2018).]
- BARANOV 2016.a. = БАРАНОВ, В. А. 2016.a. Модуль n-грамм исторического корпуса «Манускрипт»: структурные и лингвистические параметры. *Научное наследие В. А. Богородицкого и современный вектор исследований Казанской лингвистической школы. 1*. Казань: Изд-во Казан. ун-та, 50–61. URL: <https://drive.google.com/file/d/0BwBejXXryRcROHc5T3YxTENhLUk/view> (15.5.2018). [BARANOV, V. A. 2016.a. Modul' n-gramm istoričeskogo korpusa »Manuskript«: strukturnye i lingvističeskie parametry. *Naučnoe nasledie V. A. Bogorodickogo i sovremennyyj vektor issledovanij Kazanskoj lingvističeskoj školy. 1*. Kazan': Izd-vo Kazan. un-ta, 50–61. URL: <https://drive.google.com/file/d/0BwBejXXryRcROHc5T3YxTENhLUk/view> (15.5.2018).]
- BARANOV 2016.b. = БАРАНОВ, В. А. 2016.b. Опыт создания модуля n-грамм системы «Манускрипт» и оценки эффективности его использования для поиска коллокаций в корпусе М. В. Ломоносова. *Интеллектуальные системы в производстве* 4 (31): 124–131. URL: <https://drive.google.com/file/d/0BwBejXXryRcRV3ZpemZKcC1CMTQ/view> (15.5.2018). [BARANOV, V. A. 2016.b. Opyt sozdaniâ modulâ n-gramm sistemy »Manuskript« i ocenki éffektivnosti ego ispol'zovaniâ dlâ poiska kollokacij v korpuse M. V. Lomonosova. *Intellektual'nye sistemy v proizvodstve* 4 (31): 124–131. URL: <https://drive.google.com/file/d/0BwBejXXryRcRV3ZpemZKcC1CMTQ/view> (15.5.2018).]
- BARANOV 2017. = БАРАНОВ, В. А. 2017. Инструменты и методика поиска коллокаций в историческом корпусе «Манускрипт» (на примере глаголических рукописей). *Filologija* 68: 17–49. [BARANOV, V. A. 2017. Instrumenty i metodika poiska kollokacij v istoričeskom korpuse »Manuskript« (na primere glagoličeskikh rukopisej). *Filologija* 68: 17–49.]

- EVERT, S. 2004. Association Measures. *Computational Approaches to Collocations*. URL: <http://collocations.de/AM/index.html> (15.5.2018).
- НОНЛОВА 2008. = ХОХЛОВА, М. В. 2008. Экспериментальная проверка методов выделения коллокаций. *Slavica Helsingensia 34. Инструментарий русистики: Корпусные подходы*. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой (ред.), 343–357. URL: <http://www.helsinki.fi/slavicahelsingensia/preview/sh34/pdf/21.pdf> (15.5.2018). [НОНЛОВА, М. В. 2008. Ëksperimental'naâ proverka metodov vydeleniâ kollokacij. *Slavica Helsingensia 34. Instrumentarij rusistiki: Korpusnye podhody*. А. Mustajoki, М. V. Kopoteva, L. A. Birûlina, E. Û. Protasovoj (red.), 343–357. URL: <http://www.helsinki.fi/slavicahelsingensia/preview/sh34/pdf/21.pdf> (15.5.2018).]
- КОСЕТКОВА. = КОЧЕТКОВА, Н. А. Статистические языковые методы. Коллокации и коллигации. *Cyberleninka.ru*. URL: <http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii> (15.5.2018). [КОСЕТКОВА, Н. А. Statističeskie âzykovye metody. Kollokacii i kolligacii. *Cyberleninka.ru*. URL: <http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii> (15.5.2018).]
- ЛУКАШЕВИЧ; LOGAČEV 2010. = ЛУКАШЕВИЧ, Н. В.; Ю. М. ЛОГАЧЕВ. 2010. Комбинирование признаков для автоматического извлечения терминов. *Вычислительные методы и программирование 2*: 108–116. URL: [https://elibrary.ru/download/elibrary\\_15272886\\_42034432.pdf](https://elibrary.ru/download/elibrary_15272886_42034432.pdf) (15.5.2018). [ЛУКАШЕВИЧ, Н. В.; Û. М. LOGAČEV. 2010. Kombinirovanie priznakov dlâ avtomatičeskogo izvlečeniâ terminov. *Vyčislitel'nye metody i programmirovanie 2*: 108–116. URL: [https://elibrary.ru/download/elibrary\\_15272886\\_42034432.pdf](https://elibrary.ru/download/elibrary_15272886_42034432.pdf) (15.5.2018).]
- ONAL 2017. = ОНАЛ, И. О. 2017. «Коллокация» и «коллигация»: динамика развития значений. *Язык и культура: Сборник материалов XXX Международной научно-практической конференции* (С. С. Чернов (ред.), 33–39. URL: [https://elibrary.ru/download/elibrary\\_30001229\\_26220986.pdf](https://elibrary.ru/download/elibrary_30001229_26220986.pdf) (15.5.2018). [ONAL, I. O. 2017. »Kollokaciâ« i »kolligaciâ«: dinamika razvitiâ značenij. *Âzyk i kul'tura: Sbornik materialov XXX Meždunarodnoj naučno-praktičeskoj konferencii* (S. S. Černov (red.), 33–39. URL: [https://elibrary.ru/download/elibrary\\_30001229\\_26220986.pdf](https://elibrary.ru/download/elibrary_30001229_26220986.pdf) (15.5.2018).]
- PECINA, P.; P. SCHLESINGER. 2006. Combining association measures for collocation extraction. *Annual Meeting of the Association for Computational Linguistics (ACL 2006)*. Sydney: ACM, 651–658.
- VLAVACKAÂ 2011. = ВЛАВАЦКАЯ, М. В. 2011. Система базовых понятий комбинаторной лингвистики. *Мир науки, культуры, образования 1*: 32–38. URL: [https://elibrary.ru/download/elibrary\\_17783474\\_54708535.pdf](https://elibrary.ru/download/elibrary_17783474_54708535.pdf) (15.5.2018). [VLAVACKAÂ, М. В. 2011. Sistema bazovyh ponâtij kombinatornoj lingvistiki. *Mir nauki, kul'tury, obrazovaniâ 1*: 32–38. URL: [https://elibrary.ru/download/elibrary\\_17783474\\_54708535.pdf](https://elibrary.ru/download/elibrary_17783474_54708535.pdf) (15.5.2018).]
- ZALESSKAÂ 2014. = ЗАЛЕССКАЯ, В. В. 2014. Программа выявления в тексте двучленных статистически значимых осмысленных коллокаций (на материале

- русского языка). *Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS–2014)*. Санкт-Петербург, 283–289. URL: <http://ojs.ifmo.ru/index.php/IMS/article/viewFile/267/263> (15.5.2018). [ZALESSKAĀ, V. V. 2014. Programma vyāvleniā v tekste dvoučlennykh statističeski značimyh osmyslennykh kollokacij (na materiale russkogo āzyka). *Tehnologii informacionnogo obšestva v nauke, obrazovanii i kul'ture: sbornik naučnykh statej. Trudy XVII Vserossijskoj ob"edinennoj konferencii «Internet i sovremennoe obšestvo» (IMS–2014)*. Sankt-Peterburg, 283–289. URL: <http://ojs.ifmo.ru/index.php/IMS/article/viewFile/267/263> (15.5.2018).]
- ZAHAROV; HOHLOVA 2014. = ЗАХАРОВ, В. П.; М. В. ХОХЛОВА. 2014. Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации. *Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS–2014)*. Санкт-Петербург, 290–293. URL: <http://ojs.ifmo.ru/index.php/IMS/article/viewFile/268/264> (15.5.2018). [ZAHAROV, V. P.; M. V. HOHLOVA. 2014. Vydelenie terminologičeskikh slovosočetanij iz special'nyh tekstov na osnove različnykh mer associacii. *Tehnologii informacionnogo obšestva v nauke, obrazovanii i kul'ture: sbornik naučnykh statej. Trudy XVII Vserossijskoj ob"edinennoj konferencii «Internet i sovremennoe obšestvo» (IMS–2014)*. Sankt-Peterburg, 290–293. URL: <http://ojs.ifmo.ru/index.php/IMS/article/viewFile/268/264> (15.5.2018).]
- ZHANG, Z.; J. IRIA; CH. BREWSTER; F. CIRAVEGNA. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech.

## Приложение 1

Таблица 1. Триграммы (словоформы) глаголической коллекции, встречающиеся в нескольких источниках<sup>11</sup>

## Апнех 1

Table 1. Trigrams (word forms) of the glagolitic collection found in several manuscripts\*

## Prilog 1

Tablica 1. Trigrami (oblici riječi) glagoljske zbirke pronađeni u nekoliko izvora\*

| № <sub>1</sub> | № <sub>2</sub> | F | F (w1)...F (wx)                   | MI      | Адрес   |
|----------------|----------------|---|-----------------------------------|---------|---|
| 1              | 95             | 2 | енѣтъ (3) свѣтъ (33) лица (19)    | 6.56702 | ЛитС: 49.2.1-10.3 ЛитСФр: 3.2.1-17.12   |
| 2              | 133            | 2 | ѣа (6) ба (17) нашѣго (33)        | 5.72748 | ЛитС: 17.1.1-12.1 ПсДС: 28.1.1-7.19   |
| 3              | 239            | 2 | въ (147) дньсьни (2) днь (38)     | 3.99670 | ЛитС: 38.2.1-19.30 ЛитСФр: 1.2.1-24.21  |
| 4              | 266            | 2 | ѣго (1179) рави (6) недостѣни (2) | 3.65599 | ЛитСФр: 2.1.1-3.17 ЕвАс: 66.1.1-23.11   |
| 5              | 493            | 2 | по (325) въсеи (5) зѣни (34)      | 1.69062 | ПсДС: 10.1.1-1.14 ЕвАс: 147.1.1-11.32   |
| 6              | 618            | 9 | ѣ (95) бже (71) нашъ (67)         | 0.82854 | ЛитС: 14.1.1-25.10, 19.1.1-10.10, 27.1.1-7.11, 27.1.1-16.22, 59.2.1-11.13, 62.1.1-10.1, 64.1.1-22.1, 65.1.1-12.23 ЛитСФр: 2.1.1-2.1 |

<sup>11</sup> №<sub>1</sub> – номер по порядку, №<sub>2</sub> – номер в перечне триграмм выборки, F – количество триграммы в коллекции, F (w1)...F (wx) – триграмма и количество каждого из компонентов в коллекции, MI – значение меры, Адрес – адрес триграммы в рукописях, где 49 – лист, 2 – оборотная страница, .1 – номер столбца, -10 – номер строки, .3 – номер первого символа в строке.

\* №<sub>1</sub> – number in order, №<sub>2</sub> – number in the list of trigrams of the sample, F – number of trigrams in the collection, F (w1) ... F (wx) – trigram and number of each of the components in the collection, MI – measure value, Address – address trigrams in manuscripts, where 49 is a sheet, 2 is a back page, .1 is a column number, -10 is a line number, .3 is the number of the first character in a line.

\* №<sub>1</sub> – broj po redosljedu, №<sub>2</sub> – broj na popisu uzoraka trigrama, F – količina trigrama u zbirci, F(w1)...F(wx) – trigram i količina svake komponente u zbirci, MI – vrijednost mjere, address – adresa trigrama u rukopisima, gdje je 49 – list, 2 – recto stranica, .1 – broj stupca, -10 – broj retka, .3 – broj prvog znaka u retku.

| № <sub>1</sub> | № <sub>2</sub> | F  | F (wI)...F (wx)                        | MI       | Адрес   |
|----------------|----------------|----|--|----------|---|
| 7              | 824            | 2  | сѣ (285) лѣнолѣж (8) еси (106)         | -0.43844 | ПеДС: 32.1.1-13.18 ЕвАс: 71.1.1-18.1  |
| 8              | 951            | 4  | рѣба (56) тѣоѣго (161) сѣго (89)       | -1.16970 | ЛитС: 27.1.1-17.27, 41.2.1-8.13, 42.1.1-13.15 ПеДС: 2.1.1-11.10   |
| 9              | 1074           | 2  | бѣнидѣж (7) бѣ (2108) дѣиѣж (45)       | -1.89656 | ПеДС: 6.1.1-15.15 ЕвАс: 42.2.1-12.16  |
| 10             | 1167           | 2  | зѣноѣсѣтѣж (3) о (1122) тѣѣж (285)     | -2.42733 | ПеДС: 98.2.1-20.16 ЕвАс: 140.2.1-4.21   |
| 11             | 1215           | 4  | о (1122) дѣсѣлѣж (12) лѣне (169)       | -2.67340 | ПеДС: 19.1.1-14.18 ЕвАс: 49.1.2-28.1, 79.1.1-25.1, 116.1.1-23.3   |
| 12             | 1295           | 4  | бѣ (2108) вѣсѣ (50) лѣиѣж (31)         | -3.19541 | ЛитС: 56.1.1-13.28 ЕвАс: 99.1.1-5.3, 99.1.1-7.7, 154.1.1-16.26  |
| 13             | 1430           | 2  | прѣдѣлѣж (10) вѣдѣтѣж (128) бѣж (2108) | -3.91928 | ПеДС: 14.2.1-12.10 ЕвАс: 78.1.2-7.7   |
| 14             | 1459           | 29 | оѣж (83) жѣ (1185) рѣне (449)          | -4.09400 | ЛитС: 43.2.1-19.10 ЕвАс: 7.1.2-13.1, 24.2.2-21.11, 25.1.1-10.13, 26.1.1-10.10, 41.1.1-21.13, 42.1.1-16.22, 44.1.1-20.10, 45.2.1-4.11, 53.2.2-16.10, 55.2.1-26.13, 56.2.1-16.11, 58.1.1-28.1, 58.2.2-27.11, 59.2.1-8.10, 59.2.2-6.11, 62.2.1-28.11, 64.1.2-26.1, 64.2.2-5.11, 65.2.1-13.11, 70.2.2-7.12, 71.1.1-15.1, 88.1.1-15.12, 116.2.2-6.10, 119.2.1-16.24, 119.2.1-26.32, 128.1.1-25.11, 139.1.1-6.11, 156.2.1-26.29 |
| 15             | 1462           | 2  | на (1053) глѣѣж (4) вѣо (729)          | -4.10576 | ПеДС: 28.2.1-4.12 ЕвАс: 155.2.1-22.11   |
| 16             | 1468           | 2  | ѣѣо (1179) бѣжѣлѣлѣж (6) еѣтѣж (442)   | -4.13191 | ПеДС: 133.2.1-11.1 ЕвАс: 149.1.1-11.20  |
| 17             | 1482           | 2  | бѣ (2108) вѣѣлѣлѣ (38) сѣѣ (41)        | -4.20284 | ПеДС: 1.2.1-11.17 ЕвАс: 148.2.1-19.12   |
| 18             | 1557           | 3  | бѣ (2108) дѣиѣж (45) тѣѣи (72)         | -4.67417 | ПеДС: 6.1.1-15.15 ЕвАс: 14.2.2-1.1, 120.1.1-2.16  |
| 19             | 1606           | 2  | еѣтѣж (442) дѣѣлѣлѣ (6) бѣж (2108)     | -4.97022 | ПеДС: 119.1.1-1.14 ЕвАс: 47.1.1-5.13  |
| 20             | 1608           | 2  | дѣ (730) не (1283) вѣжѣрѣлѣж (6)       | -4.97772 | ПеДС: 83.1.1-1.1 ЕвАс: 83.2.1-15.10   |
| 21             | 1609           | 2  | и (1348) бѣжѣлѣ (4) на (1053)          | -4.99259 | ЛитС: 38.2.1-17.19 ПеДС: 22.2.1-8.1   |

| № <sub>1</sub> | № <sub>2</sub> | F | F (w1)...F (wx)                    | MI       | Адрес  |
|----------------|----------------|---|------------------------------------|----------|--|
| 22             | 1652           | 2 | по (325) ерѣдѣ (18) же (1185)      | -5.28059 | ПеДС: 91.1.1-14.15 ЕвАс: 118.1.1-8.13  |
| 23             | 1694           | 3 | сепо (89) рѣдѣ (106) не (1283)     | -5.49969 | ПеДС: 57.2.1-23.16 ЕвАс: 27.2.1-17.12, 62.1.1-17.11                          |
| 24             | 1769           | 4 | и (4551) по (325) тоидѣ (16)       | -6.05196 | ЛитС: 44.2.1-6.11 ЕвАс: 54.1.1-4.12, 66.1.1-12.14, 124.2.1-23.14             |
| 25             | 1775           | 2 | вѣдѣ (128) вѣ (2108) рѣдѣ (45)     | -6.08921 | ПеДС: 14.2.1-12.10 ЕвАс: 78.1.2-8.10   |
| 26             | 1800           | 2 | прѣдѣ (137) тобож (22) и (4551)    | -6.26513 | ПеДС: 6.1.1-1.13 ЕвАс: 62.2.2-24.3   |
| 27             | 1817           | 2 | не (1283) вѣзвратитѣ (6) сѣ (1975) | -6.41360 | ПеДС: 83.1.1-1.15 ЕвАс: 83.2.1-15.10   |
| 28             | 1836           | 2 | не (1283) вѣдѣ (11) вѣко (1179)    | -6.54378 | ПеДС: 41.2.1-21.1 ЕвАс: 74.1.1-12.7  |
| 29             | 1896           | 3 | сѣдѣ (22) вѣ (729) вѣ (2108)       | -6.98160 | ЛитС: 43.2.1-9.12 ПеДС: 48.1.1-19.17, 96.2.1-18.18                           |
| 30             | 1913           | 2 | сѣ (1975) вѣ (2108) чрѣдѣ (6)      | -7.12996 | МинС: 1.2.1-2.16 ЕвАс: 136.1.1-4.11  |
| 31             | 1929           | 3 | сѣ (1975) вѣ (533) нево (39)       | -7.26177 | ПеДС: 114.1.1-19.10 ЕвАс: 106.1.2-16.10, 110.2.1-27.3                        |
| 32             | 1986           | 2 | вѣко (1179) тако (72) ерѣ (442)    | -7.71687 | ПеДС: 72.2.1-16.15 ЕвАс: 28.2.1-6.12   |
| 33             | 1997           | 2 | вѣ (729) на (1053) настѣ (52)      | -7.80620 | ПеДС: 118.1.1-13.19 ЕвАс: 108.1.2-21.1                                       |
| 34             | 2020           | 5 | сѣ (1975) о (1122) нево (50)       | -7.95712 | ПеДС: 73.1.1-16.10, 80.2.1-5.17 ЕвАс: 92.2.1-1.11, 92.2.1-3.11, 152.2.1-8.13 |
| 35             | 2034           | 2 | и (4551) вѣ (533) вѣдѣ (23)        | -8.28922 | ЕвАс: 3.2.1-16.1 КГЛ: 1.1.1-25.12  |
| 36             | 2101           | 3 | и (4551) на (1053) зѣли (34)       | -9.25045 | ЛитС: 54.2.1-20.6 ЕвАс: 71.1.2-26.11, 133.1.1-19.11                          |
| 37             | 2109           | 2 | дѣ (730) не (1283) вѣдѣ (128)      | -9.39275 | ПеДС: 90.2.1-18.1 ЕвАс: 83.2.1-23.13   |
| 38             | 2111           | 2 | вѣ (2108) дѣ (28) вѣ (2108)        | -9.44637 | ПеДС: 112.1.1-5.31 ЕвАс: 42.2.2-2.5  |
| 39             | 2122           | 3 | и (4551) присно (10) и (4551)      | -9.59660 | ЛитС: 19.1.1-18.11, 51.2.1-13.11 ПеДС: 2.1.1-12.13                           |
| 40             | 2134           | 2 | сѣ (1975) вѣ (2108) вѣдѣ (38)      | -9.79292 | ПеДС: 47.2.1-24.12 ЕвАс: 148.2.1-19.3  |

| № <sub>1</sub> | № <sub>2</sub> | F | F (wI)...F (wx)                 | MI        | Адрес  |
|----------------|----------------|---|---------------------------------|-----------|--|
| 41             | 2159           | 6 | с.А (1975) ◊ (1122) ТРЕНЬ (285) | -10.20505 | ПеДС: 11.1.1-4.6, 52.2.1-2.10, 78.1.1-13.1, 93.1.1-17.16, 94.1.1-23.15 ЕвАс: 89.2.1-19.3 |
| 42             | 2178           | 2 | с.А (1975) НА (1053) НТЬ (186)  | -11.08278 | ПеДС: 70.2.1-7.10 ЕвАс: 152.2.1-22.14  |
| 43             | 2195           | 2 | ЖЕ (1185) РЕЧЕ (449) НЕ (1283)  | -11.90225 | ЛитС: 43.2.1-19.13 ЕвАс: 58.1.1-28.12  |
| 44             | 2219           | 2 | А (1001) НЕ (1283) БТЬ (2108)   | -13.88989 | ПеДС: 120.2.1-12.5 ЕвАс: 61.1.1-3.13   |
| 45             | 2220           | 2 | ЛА (529) И (4551) НЕ (1283)     | -14.08009 | ПеДС: 20.1.1-4.10 ЕвАс: 13.1.2-6.10  |
| 46             | 2221           | 2 | И (4551) ,АА (730) НЕ (1283)    | -14.54472 | ПеДС: 140.2.1-12.11 ЕвАс: 56.2.1-20.10   |
| 47             | 2225           | 3 | с.А (1975) И (4551) НЕ (1283)   | -15.39564 | ПеДС: 5.1.1-2.18 ЕвАс: 72.2.1-22.1, 73.2.1-10.1  |

## Приложение 2

Группы из первых 250 биграмм, построенных на основе словоформ

## Appendix 2

Groups of the first 250 bigrams built on the basis of word forms

## Dodatak 2

Grupe od prvih 250 bigrama, oblikovanih na temelju oblika riječi

ПРЕДЛОГ + СУЩЕСТВИТЕЛЬНОЕ (55 – 360×):<sup>12</sup> *вс* (9) *порака* (2) 2×,<sup>13</sup> *из* (27) *оуѣтъ* (5) 2×, *подъ* (24) *крикъ* (6) 2×, *до* (139) *коньца* (7) 7×, *до* (139) *скота* (2) 2×, *прѣдъ* (137) *очима* (9) 7×, *по* (325) *срѣдѣ* (18) 17× и др., наибольшее значение меры – **13,274**,<sup>14</sup> наименьшее – 5,140;

СУЩЕСТВИТЕЛЬНОЕ + ПРИТЯЖАТЕЛЬНОЕ МЕСТОИМЕНИЕ (49 – 252×): *мои*: *очима* (6) *моима* (2) 2×, *срдѣце* (8) *моѣ* (155) 8×, *оца* (4) *моѣго* (113) 3× и др., *нашь*: *грѣхъ* (26) *нашихъ* (20) 5×, *бже* (71) *нашъ* (67) 20×, *ба* (17) *нашего* (33) 2× и др., *свои*: *люди* (26) *своѣ* (80) 11×, *село* (13) *своѣ* (41) 2×, *лице* (21) *своѣ* (41) 3× и др., *твои*: *милость* (7) *твоѣ* (24) 2×, *дхъ* (3) *твон* (72) 2×, *братъ* (11) *твон* (72) 5×, *дѣсница* (3) *твоѣ* (157) 3× и др., *его*: *главѣ* (4) *его* (729) 2×, *нозѣ* (5) *его* (729) 2× и др., значения меры – **13,859-5,349**;

ПРЕДЛОГ + МЕСТОИМЕНИЕ (24 – 186×): *ис* (8) *тебѣ* (74) 2×, *прѣдъ* (137) *твоѣ* (22) 10×, *прѣдъ* (137) *нигъ* (36) 11×, *съ* (285) *мноѣ* (11) 5× и др., значения меры – 8,235-5,340;

ОБОРОТ С *яко* (3 – 92×): *ѣко* (1179) *каплѣ* (2) 2×, *ѣко* (1179) *мрътвѣ* (3) 2×, *ѣко* (1179) *влагъ* (33) 18×, значения меры – 6,241-5,367;

ГЛАГОЛ + *ся* (21 – 83×): *оврѣтѣ* (16) *сѣ* (79) 2×, *веселѣ* (2) *сѣ* (1975) 2×, *обратитѣ* (2) *сѣ* (1975) 2× и др., значения меры – 7,141-5,192;

МЕСТОИМЕНИЕ + ЧАСТИЦА *же* (2 – 65×): *онъ* (83) *же* (1185) (61×), *тѣмъ* (8) *же* (1185) (4×), значения меры – 5,789-5,234;

МЕСТОИМЕНИЕ + ПОСЛЕЛОГ (5 – 57×): *сеѣго* (89) *ради* (106) 34×, *сеѣго* (89) *ради* (43) 9×, *чѣсо* (20) *ради* (43) 2× и др., значения меры – 8,238-5,707;

<sup>12</sup> Где 55 – количество сочетаний, 360 – количество случаев встречаемости в текстах коллекции.

<sup>13</sup> На первом месте приводятся биграмы, имеющие наибольшее статистическое значение.

<sup>14</sup> Выделены наибольшие для выборки значения меры.

ПРЕДИКАТИВНЫЕ КОНСТРУКЦИИ (7 – 38×): *свѣтъ* (3) *свѣтъ* (33) 2×, *тъ* (68) *еси* (106) 5×, *если* (95) *азъ* (117) 15× и др., значения меры – **10,815-5,195**;

СОЧЕТАНИЕ *Господь Богъ* (3 – 31×): *ѣа* (6) *ѣа* (17) 3×, *ѣи* (338) *ѣже* (75) 17×, *ѣъ* (193) *ѣтъ* (139) 11×, значения меры – **11,357-5,192**;

ЧАСТИЦА *не* + ГЛАГОЛ (7 – 31×): *не* (1283) *възмогътъ* (2), *не* (1283) *могътъ* (3), *не* (1283) *отъврати* (5) и др., значения меры – 6,119-5,312;

ГЛАГОЛЬНО-ПРЕДЛОЖНЫЕ ФОРМЫ (11 – 30×): *нападѣ* (3) *на* (1053) 3×, *нападѣ* (4) *на* (1053) 3×, *сѣдѣ* (3) *на* (1053) 2, *заповѣсть* (3) *о* (1122) 2× и др., значения меры – 6,404-5,403;

ГЛАГОЛ / ПРИЧАСТИЕ + МЕСТОИМЕНИЕ (9 – 24×): *оуготованое* (2) *вамъ* (183) 2×, *помози* (6) *намъ* (120) 3×, *даждѣ* (19) *ми* (60) 3× и др., значения меры – 8,929-5,705;

ФОРМЫ МЕСТОИМЕНИЯ *весь* + СУЩЕСТВИТЕЛЬНОЕ (6 – 24×): *всему* (3) *мироу* (7) и др. 18×, *въсеи* (5) *земли* (34) 2×, *въси* (79) *людие* (26) 4×, значения меры – **13,052-7,440**;

ПРИЛАГАТЕЛЬНОЕ / ПРИЧАСТИЕ + *быти* (3 – 13×): *живи* (2) *вждете* (26) 2×, *прѣданъ* (10) *вждетъ* (128) 6×, *живъ* (14) *вждетъ* (128) 5×, значения меры – **11,744-7,959**;

ПЕРФЕКТНЫЕ ФОРМЫ ГЛАГОЛОВ (3 – 11×): *далъ* (29) *еси* (106) 5×, *възвеличилъ* (6) *естъ* (442) 4×, *еси* (106) *далъ* (29) 2×, значения меры – 7,180-5,858;

ЧАСТИЦА *да* + ГЛАГОЛ (4 – 9×): *да* (730) *омочитъ* (2) 2×, *да* (730) *просвѣтитъ* (3) 2× и др., значения меры – 6,933-5,196;

СУЩЕСТВИТЕЛЬНОЕ + ПРИЛАГАТЕЛЬНОЕ (3 – 7×): *рави* (6) *недостонни* (2) 2×, *дньсьни* (2) *днь* (38) 2×, *вѣчьны* (10) *ѣже* (71) 3×, значения меры – **13,859-5,558**;

ГЛАГОЛ + СУЩЕСТВИТЕЛЬНОЕ (2 – 4×): *сѣжежетъ* (2) *огнеи* (6) 2×, *простѣрък* (3) *рѣкѣ* (29) 2×, значения меры – **13,859-11,001**;

СОЧЕТАНИЕ *другъ друга* (1 – 4×): *дроугъ* (8) *дроуга* (13) 4×, значение меры – **11,744**;

СОЧЕТАНИЕ *свѣтъ лица* (1 – 3×): *свѣтъ* (33) *лица* (19) 3×, значение меры – 8,737 и нек. др.

Приложение 3  
Таблица 2. Триграммы (леммы) глаголической коллекции, встречающиеся в нескольких источниках

Appendix 3  
Table 2. Trigrams (lemmas) of the glagolitic collection, found in several manuscripts

Prilog 3  
Tablica 2. Trigrami (leme) glagoljske zbirke, koji se nalaze u više izvora

| № <sub>1</sub> | № <sub>2</sub> | F | F (w1)...F (wx)                            | MI       | Адрес и текстовая форма   |
|----------------|----------------|---|--|----------|---|
| 1              | 70             | 2 | иже (1087) насадици (8)<br>десница (33)    | -0.68623 | ЛитСФр: иже насади десница 1.2.1-10.26 ПсДС: иже насади десница 90.1.1-9.12   |
| 2              | 89             | 3 | пладъ (40) чръво (16)<br>твон (1035)       | -1.30808 | ПсДС: пладъ чръво твоего 128.2.1-17.1, пладъ чръво твоего 130.1.1-25.27<br>ЕвАс: пладъ чръво твоего 144.2.1-15.10   |
| 3              | 113            | 2 | око (90) лон (1140)<br>сльза (6)           | -1.78733 | ЛитСФр: очина лонна слъзы 1.1.1-8.18 ПсДС: очина лонна слъзы 140.2.1-5.15   |
| 4              | 133            | 5 | възвратици (52) сѧ<br>(2950) въсплати (13) | -2.16115 | ПсДС: възвратици сѧ въсплатъ 78.1.1-7.13, възврати сѧ въсплатъ 116.2.1-7.13, възврати сѧ въсплатъ 116.2.1-12.14, възвратици сѧ въсплатъ 129.1.1-23.20 ЕвАс: възвратици сѧ въсплатъ 83.2.1-15.10 |
| 5              | 168            | 2 | сльныце (13) и (15437)<br>лоуна (7)        | -2.97760 | ПсДС: сльныце і лѧна 143.1.1-17.15 ЕвАс: сльныци і лоуна 71.1.2-24.10   |
| 6              | 171            | 3 | и (15437) коль (20)<br>красьна (7)         | -3.01412 | ЛитС: и коль красно 41.2.1-16.19 ПсДС: і коль красно 130.2.1-17.27<br>ЕвАс: и коль красно 125.2.1001-1.16   |
| 7              | 201            | 2 | въстоукъ (13) и (15437)<br>зандаць (10)    | -3.49217 | ПсДС: въстоукъ і зандаць 111.1.1-23.24 ЕвАс: въстоукъ і зандаць 63.1.1-18.5   |
| 8              | 227            | 2 | оутвердици (37) сила<br>(66) твон (1035)   | -3.82496 | ПсДС: ѡтверди слово твоего 82.2.1-11.10 КГЛ: сила твон оутверди 4.2.1-5.16  |

| № <sub>1</sub> | № <sub>2</sub> | F  | F (w1)...F (wx)                          | MI       | Адрес и текстовая форма  |
|----------------|----------------|----|--|----------|--|
| 9              | 260            | 3  | слово (142) твой (1035)<br>истина (36)   | -4.30582 | ЛитСФр: слово твоей истины 3.2.1-20.10 ПеДС: словесъ твоюхъ истина 125.2.1-11.14 ЕвАС: слово твоѣ истина 99.1.1-1.11   |
| 10             | 327            | 2  | преславити (33) имя<br>(164) твой (1035) | -4.97306 | ПеДС: преслаблать имя твоѣ 93.2.1-24.12 ЕвАС: преслави имя твоѣ 27.1.1-7.15  |
| 11             | 356            | 15 | родъ (58) и<br>родъ (58)                 | -5.27887 | ЛитС: родъ и родъ 52.2.1-13.12 ПеДС: родъ и родъ 42.2.1-12.3, рода и рода 79.2.1-23.3, родъ и родъ 89.2.1-1.3, родъ и родъ 95.1.1-23.32, родъ и родъ 95.2.1-6.3, рода и рода 103.2.1-10.13, родъ и родъ 104.2.1-10.11, родъ и родъ 105.1.1-10.1, родъ и родъ 110.2.1-9.20, родъ и родъ 122.2.1-22.10, родъ и родъ 131.2.1-14.11, родъ и родъ 139.1.1-21.1, родъ и родъ 139.2.1-23.11, родъ и родъ 142.1.1-12.1 |
| 12             | 406            | 2  | свои (577) заповѣдѣти<br>(16) ѿ (1050)   | -5.76430 | ПеДС: своиъ заповѣтъ ѿ 98.2.1-20.10 ЕвАС: своиъ заповѣтъ ѿ 140.2.1-4.14  |
| 13             | 408            | 2  | гора (90) и (15437) хль-<br>ль (7)       | -5.76901 | ПеДС: горы и хльви 143.2.1-5.1 ЕвАС: гора и хльви 138.2.1-7.27   |
| 14             | 425            | 2  | и (15437) за<br>оудильноженіе (4)        | -5.92923 | ЛитСФр: и за оудильноженіи 2.1.1-6.16 ЕвАС: и за оудильноженіи 49.1.1-5.13   |
| 15             | 430            | 10 | въ (4705) въсь (86)<br>миръ (138)        | -5.96854 | ЛитС: въ весь миръ 56.1.1-13.28, въ весь миръ 61.1.1-24.10 ЕвАС: въ весь миръ 80.2.2-5.7, въ весь миръ 97.2.1-17.11, въ весь миръ 99.1.1-5.3, въ весь миръ 99.1.1-7.7, въ весь миръ 102.1.1-28.1, въ весь миръ 117.1.1-15.1, въ весь миръ 117.1.1-18.27, въ весь миръ 154.1.1-16.26  |
| 16             | 442            | 2  | иъмиѣ (52) и<br>присно (15)              | -6.07713 | ЛитС: иъмиѣ и присно 51.2.1-13.11 ПеДС: иъмиѣ и присно 2.1.1-12.2  |

| № <sub>1</sub> | № <sub>2</sub> | F  | F(w1)...F(wx)                         | MI       | Адрес и текстовая форма   |
|----------------|----------------|----|---------------------------------------|----------|---|
| 17             | 443            | 30 | онгъ (289) же (1178)<br>реци (534)    | -6.08655 | ЛигС: онгъ же рече 43.2.1-19.10 ЕвАс: онгъ же рече 7.1.2-13.1, онгъ же рече 24.2.2-21.11, онгъ же рече 25.1.1-10.13, онгъ же рече 26.1.1-10.10, онгъ же рече 41.1.1-21.13, онгъ же рече 42.1.1-16.22, онгъ же рече 44.1.1-20.10, онгъ же рече 45.2.1-4.11, онгъ же рече 53.2.2-16.10, онгъ же рече 55.2.1-26.13, онгъ же рече 56.2.1-16.11, онгъ же рече 58.1.1-28.1, онгъ же рече 58.2.2-27.11, онгъ же рече 59.2.1-8.10, онгъ же рече 59.2.2-6.11, онгъ же рече 62.2.1-28.11, онгъ же рече 64.1.2-26.1, онгъ же рече 64.2.2-5.11, онгъ же рече 65.2.1-13.11, онгъ же рече 70.2.2-7.12, онгъ же рече 71.1.1-15.1, онгъ же рече 88.1.1-15.12, онгъ же рече 116.2.2-6.10, онгъ же рече 119.2.1-16.24, онгъ же рече 119.2.1-26.32, онгъ же рече 128.1.1-25.11, онгъ же рече 139.1.1-6.11, она же рече 153.1.1-15.23, онгъ же рече 156.2.1-26.29 |
| 18             | 452            | 2  | заповѣдѣти (16) о (1050) ты (752)     | -6.14646 | ПеДС: заповѣтѣрь о тѣвѣ 98.2.1-20.16 ЕвАс: заповѣтѣрь о тѣвѣ 140.2.1-4.21   |
| 19             | 455            | 2  | вѣсь (981) дѣнь (210)<br>животѣ (62)  | -6.16225 | ЛигС: вѣсь дѣни животѣ 47.1.1-8.10 ПеДС: вѣсь дѣнь животѣ 129.1.1-9.16  |
| 20             | 458            | 3  | вѣти (1540) къ (419)<br>тоноу (30)    | -6.17715 | ПеДС: вѣтѣрь къ тоноу 82.2.1-2.1 ЕвАс: вѣтѣ къ тоноу 15.1.1-3.11, къ тоноу вѣсѣнь 98.2.1-7.1  |
| 21             | 482            | 2  | вѣти (1540) во (262)<br>дѣстоинѣ (35) | -6.30712 | ЛигСФр: во вѣтѣ дѣстоинѣ 2.1.1-7.25 ЕвАс: вѣсѣть во дѣстоинѣ 54.2.1-21.1  |
| 22             | 488            | 3  | принѣти (130) вѣсѣнѣти (57) съ (2950) | -6.35249 | МинС: принѣдѣть вѣсѣдѣтѣ съ 2.2.1-2.10 ЕвАс: принѣдѣть вѣсѣни съ 135.2.1-2.20, принѣдѣть вѣсѣни съ 140.2.1-24.22  |
| 23             | 494            | 2  | дѣша (133) рагѣ (112)<br>тѣвои (1035) | -6.43375 | ЛигС: дѣша раба тѣвогѣ 32.2.1-13.20 ПеДС: дѣшѣ раба тѣвогѣ 93.2.1-15.10   |
| 24             | 497            | 2  | дѣнѣ (138) на (1090)<br>дѣди (105)    | -6.46858 | ЛигС: дѣнѣ на дѣди 58.1.1-21.12 ПеДС: дѣнѣ на дѣди 93.1.1-21.1  |

| № <sub>1</sub> | № <sub>2</sub> | F  | F (w1)...F (wx)                           | MI       | Адрес и текстовая форма   |
|----------------|----------------|----|---|----------|---|
| 25             | 513            | 3  | исповѣданѣ (12) съ (2950) ты (752)        | -6.63678 | ПеДС: исповѣданюхъ съ текѣ 82.2.1-23.22 ЕвАс: исповѣданъ съ текѣ 60.2.1-13.17, исповѣданъ съ текѣ 129.1.1-20.31   |
| 26             | 516            | 8  | ангелствъ (125) твон (1035) господь (568) | -6.68664 | ЛитС: ꙗ ангелствъ твоѣхъ 22.2.1-10.11 ПеДС: ангелствъ твоѣхъ ꙗ 89.1.1-3.22, ꙗ ангелствъ твою 93.1.1-18.11, ангелствъ твою ꙗ 95.1.1-23.1, ангелствъ твоѣхъ ꙗ 98.1.1-12.11, ангелствъ твоѣхъ ꙗ 100.2.1-11.1, ангелствъ твоѣхъ ꙗ 121.2.1-17.1, ꙗ ангелствъ твоѣхъ 134.1.1-4.19   |
| 27             | 517            | 2  | ти (298) поклонити (21) съ (2950)         | -6.69368 | ПеДС: поклонити ти съ 73.1.1-9.12 ЕвАс: ти поклонити съ 28.2.2-18.14  |
| 28             | 518            | 21 | къ (419) томоу (30) и (15437)             | -6.69518 | ПеДС: і къ томи 51.1.1-15.10, къ томи не 51.1.1-15.11, къ томи і 94.2.1-20.28 ЕвАс: і томоу не 15.1.1-3.11, і къ томоу 15.2.1-18.10, къ томоу не 15.2.1-18.10, к томоу не 26.1.2-26.3, к томоу не 30.1.1-9.3, его к томоу 49.2.1-13.5, къ томоу не 75.2.2-6.10, і к томоу 78.1.1-3.10, і томоу не 78.1.1-3.10, къ томоу не 93.2.2-14.5, і к томоу 96.2.2-5.12, к томоу не 96.2.2-6.1, і къ томоу 97.1.1-13.4, къ томоу не 97.1.1-13.6, і къ томоу 98.2.1-6.12, его к томоу 116.1.1-11.11, і къ томоу 156.2.1-29.13, къ томоу не 156.2.1-29.15 |
| 29             | 563            | 13 | вѣсти (1540) на (1090) него (93)          | -7.07324 | ЛитСФр: ѳси на нѣсѣхъ 1.2.1-18.23 ПеДС: ѳствъ на нѣсѣи 81.2.1-14.13 ЕвАс: ѳствъ на нѣсѣхъ 32.2.2-13.3, ѳствъ на нѣсѣхъ 33.2.1-9.1, ѳствъ на нѣсѣхъ 34.1.2-4.8, ѳствъ на нѣсѣхъ 34.1.2-11.8, ѳствъ на нѣсѣхъ 47.2.1-3.10, ѳствъ на нѣсѣхъ 73.1.1-15.1, ѳствъ на нѣсѣхъ 113.2.2-21.9, ѳствъ на нѣсѣхъ 119.1.1-8.12, ѳствъ на нѣсѣхъ 123.1.1-23.22, ѳствъ на нѣсѣхъ 124.2.1-4.11, ѳствъ на нѣсѣхъ 150.1.1-14.10  |
| 30             | 566            | 5  | днь (101) и (15437) ноцъ (39)             | -7.09149 | ПеДС: днь і ноцъ 1.2.1-8.11, днь и ноцъ 2.1.1-15.10, днь и ноцъ 144.1.1-6.13 ЕвАс: днь і ноцъ 67.1.1-22.7, днь и ноцъ 142.1.1-27.14   |
| 31             | 574            | 2  | помиловати (37) азъ (1211) господь (568)  | -7.15688 | ЛитСФр: помиловіи ми ꙗ 1.1.1-7.13 ПеДС: помиловіи ми ꙗ 93.2.1-13.22   |

| № <sub>1</sub> | № <sub>2</sub> | F | F (w1)...F (wx)                           | MI       | Адрес и текстовая форма  |
|----------------|----------------|---|---|----------|--|
| 32             | 584            | 2 | петрухъ (79) и павълъ (22) (15437)        | -7.23301 | МинС: петра и павъла 2.1.1-17.14 ЕвАс: петра и павла 149.2.1-25.1  |
| 33             | 591            | 3 | принти (130) къ ты (752) (419)            | -7.25850 | ПеДС: прѣдѣтъ къ тѣвѣ 98.2.1-18.10 ЕвАс: принти къ тѣвѣ 42.1.1-15.28, прѣти къ тѣвѣ 54.2.1-26.10   |
| 34             | 592            | 3 | вели (44) по (960) въсь (981)             | -7.27517 | ПеДС: вели по въсьи 58.2.1-6.1, вели по въсьи 101.1.1-1.19 ЕвАс: вели по въсьи 147.1.1-11.26   |
| 35             | 601            | 3 | во (262) выти (1540) доухъ (109)          | -7.36106 | ПеДС: ибѣтъ по дѣа 131.2.1-21.22 ЕвАс: во въ дѣтъ 32.1.1-2.2, дѣтъ во ѣтъ 90.1.2-3.10  |
| 36             | 603            | 5 | господъ (568) когъ (441) нашъ (295)       | -7.37272 | ЛитСФр: гѣ бѣе нашъ 2.1.1-2.1 ПеДС: гѣи бѣс нашихъ 84.1.1-11.1, гѣ бѣе нашъ 107.2.1-26.21, гѣи бѣс нашихъ 127.2.1-6.3 ЕвАс: гѣ въ нашъ 115.2.1-19.11 |
| 37             | 607            | 2 | прѣстъ (11) свои (577) въ (4705)          | -7.38753 | ПеДС: въ прѣстъ свою 107.1.1-28.11 ЕвАс: прѣста своего въ 55.1.2-14.12   |
| 38             | 630            | 2 | выти (1540) оуготовати (34) прѣдѣтъ (615) | -7.49632 | ПеДС: оуготовагъ еси прѣдѣтъ 32.1.1-16.1 ЕвАс: гѣи оуготовагъ прѣдѣтъ 142.1.1-8.29   |
| 39             | 632            | 2 | сѣказаати (47) азъ (1211) господъ (568)   | -7.50202 | ПеДС: сѣказаи азъ гѣ 138.1.1-7.1 ЕвАс: сѣказаи азъ гѣ 107.1.2-23.13  |
| 40             | 638            | 3 | сѣласти (56) азъ (1211) отъ (726)         | -7.52390 | ПеДС: сѣи азъ отъ 31.1.1-4.1, сѣи азъ отъ 76.2.1-14.26 ЕвАс: сѣи азъ отъ 27.1.1-2.18   |
| 41             | 674            | 2 | са (2950) къ (419) цръ-кы (32)            | -7.79300 | ЛитСФр: са къ цръки 1.1.1-4.11 ПеДС: са къ цръки 133.2.1-8.1   |
| 42             | 686            | 3 | прѣдѣтъ (615) въсь (981) люди (105)       | -7.88753 | ПеДС: прѣдѣтъ въсьи люди 117.2.1-25.23, прѣдѣтъ въсьи люди 118.1.1-7.23 ЕвАс: прѣдѣтъ въсьи люди 57.2.2-27.3   |
| 43             | 687            | 2 | ни (165) едлингъ (356) отъ (726)          | -7.90158 | ПеДС: ни едлингъ отъ 109.2.1-24.14 ЕвАс: ни едлингъ отъ 125.1.1-3.12   |

| № <sub>1</sub> | № <sub>2</sub> | F | F (w1)...F (wx)                      | MI       | Адрес и текстовая форма   |
|----------------|----------------|---|--------------------------------------|----------|---|
| 44             | 692            | 2 | по́дъ (78) нога (36) и (15437)       | -7.92513 | ЛитС: по́дъ но́гъ ѿо 42.1.1-16.13 ПедС: по́дъ но́гма ѿо 22.2.1-7.10       |
| 45             | 705            | 2 | сѣньдѣрь (40) твои (1035) иже (1087) | -7.97918 | ПедС: сѣньдѣрь твои иже 82.1.1-5.13 ЕвАс: сѣней твои́ же 16.1.2-18.12     |
| 46             | 707            | 2 | по (960) сръдѣд (40) же (1178)       | -7.98664 | ПедС: по сръдѣд же 91.1.1-14.15 ЕвАс: по сръдѣд же 118.1.1-8.13           |
| 47             | 728            | 2 | вѣсь (981) написати (17) сѣ (2950)   | -8.10777 | ПедС: вѣсь написѣтъ сѣ 134.2.1-20.25 ЕвАс: вѣси написѣтъ сѣ 132.2.1-13.25 |
| 48             | 732            | 2 | вѣ (4705) нога (36) нашъ (295)       | -8.13017 | ПедС: ногы нашѣ вѣ 127.1.1-12.11 ЕвАс: вѣ ногѣхъ нашѣ 114.2.2-13.10       |
| 49             | 735            | 2 | и (15437) на (1090) вѣстаніе (3)     | -8.14488 | ЛитС: ѿ на вѣстаніе 56.2.1-22.15 ЕвАс: ѿ на вѣстаніе 142.1.1-15.21        |
| 50             | 767            | 2 | вѣлозкити (8) быти (1540) вѣ (4705)  | -8.34439 | ЛитС: вѣлозкитѣ ѿи вѣ 47.1.1-11.12 ЕвАс: вѣлози ѿ вѣ 11.2.2-10.14         |

## Sažetak

Viktor A. Baranov

O PITANJU PRIMJENE STATISTIČKIH METODA U PRONALAZENJU  
KOLOKACIJA I KOLIGACIJA U NAJSTARIJIM SLAVENSKIM TEKSTOVIMA  
(NA MATERIJALU GLAGOLJSKIH RUKOPISA KORPUSA »MANUSKRIFT«)

Rad je posvećen pitanjima metodologije pretraživanja i procjene stabilnosti ustaljenih kolokacija u zbirci glagoljskih tekstova povijesnog korpusa *Manuskript: slavjanskoe pis'mennoe nasledie* (*Manuskript: slavensko pismo nasleđe* – manuscripts.ru). Prikazane su mogućnosti modula n-gram za ekstrakciju kolokacija – koje se sastoje od riječi i njihovih oblika ili lema – s različitim brojem komponenata i različitom učestalošću. Analiziraju se bigrami i trigrami izdvojeni pomoću statističke mjere uzajamne obavijesti (eng. *Mutual Information* – akr. MI), koji se pojavljuju istodobno u nekoliko rukopisa ove zbirke.

Posebna se pozornost posvećuje n-gramima s visokom statističkom vrijednošću mjere MI. U skladu s obilježjima mjere, najveću vrijednost u zbirci imaju rijetke kombinacije. Analiza takvih rijetkih bigrama na temelju oblika riječi omogućila je identificiranje koherentnih gramatičkih konstrukcija – koligacija. Pokazano je da su trigrami koji se sastoje od tekstualnih oblika, ne samo gramatičke, već i semantičke cjeline – kolokacije. Bigrami s komponentama-lemama su raznoliki – prijedložno-padežne kolokacije s imenicama i povratno-posvojnim zamjenicama te ostale atributivne konstrukcije, glagolsko-imeničke odnosno konstrukcije itd. Pomoću analize ovih skupina pronađene su kako koligacije, tako i kolokacije. Najproduktivniji rezultat bilo je izdvajanje trigrama na temelju lema: većina prvih nekoliko desetaka kombinacija koje imaju maksimalnu vrijednost u skladu sa statističkom mjerom uzajamne obavijesti – gramatičke su i semantičke cjeline ili njihovi dijelovi.

Donosi se zaključak o učinkovitosti primjene statističkih metoda za ekstrakciju kolokacija i koligacija iz korpusa srednjovjekovnih slavenskih rukopisa. Navedeno je da se za složeno rješenje postavljenog zadatka trebaju koristiti različite vrste n-grama – dvokomponentne i trokomponentne, utemeljene na tekstualnim oblicima i lemama, sa slobodnim i fiksnim redoslijedom komponenata. Prisutnost gramatički i semantički cjelovitih kolokacija koje se ponavljaju u različitim rukopisima omogućuje nam zaključak o nadtekstnoj prirodi takvih kolokacija.

Ključne riječi: tekstualni korpus »Manuskript«, manuscripts.ru, glagoljski rukopis, jezična statistika, modul n-gram, kolokacija, koligacija

## Summary

Viktor A. Baranov

### ON THE QUESTION OF THE APPLICATION OF STATISTICAL METHODS TO SEARCH FOR COLLOCATIONS AND COLLIGATIONS IN OLD SLAVONIC TEXTS (IN GLAGOLITIC MANUSCRIPTS FROM THE CORPUS »MANUSCRIPT«)

The paper deals with the questions concerning the methodology used to search for fixed collocations in the collection of Glagolitic texts in the historical corpus *Manuscript: Slavic written heritage* (manuscripts.ru) and to evaluate their stability. It demonstrates the possibilities of the n-gram module to extract collocations, consisting of words and their textual forms or lemmas, with different numbers of components and different frequency of occurrence. Analyzed are digrams and trigrams extracted using the statistical measure of *Mutual Information* (acr. MI) that occur simultaneously in several manuscripts from the collection.

Particular attention is given to n-grams with high statistical MI values. In accordance with the specifics of the measure, the greatest values belong to the collocations that are rare in the collection. The analysis of such digrams based on the word forms has enabled an identification of coherent grammatical structures – colligations. Trigrams consisting of textual forms are shown to be not only grammatical, but also semantic units – collocations. Digrams with components-lemmas have different forms: preposition-noun collocations, preposition-possessive pronoun collocations and other attributive constructions, relative verb-noun constructions, etc. The analysis of these groups identified both colligations and collocations. Extraction of trigrams on the basis of lemmas was the most productive – the greatest part of the first few dozens of collocations with a maximum MI value are grammatical and semantic units or their parts. A conclusion is made about the efficiency of application of statistical methods for the extraction of collocations and colligations from the corpora comprising medieval Slavonic manuscripts. A complex solution of the given problem requires the use of different types of n-grams – two-components and triple-components, based on textual forms and lemmas, with free and fixed component order. The presence of grammatical and semantic units repeated in various manuscripts leads to a conclusion about the supra-textual nature of such collocations.

**Keywords:** textual corpus »Manuscript, manuscripts.ru, Glagolitic manuscript, linguistic statistics, n-gram module, collocation, colligation

Viktor A. Baranov  
Department of Linguistics  
Izhevsk State Technical University after M.T. Kalashnikov  
Izhevsk (Russia)  
victor.a.baranov@gmail.com