КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 805.3:004

В. А. Баранов, доктор филологических наук, профессор
ИжГТУ имени М. Т. Калашникова

ОПЫТ СОЗДАНИЯ МОДУЛЯ N-ГРАММ СИСТЕМЫ «МАНУСКРИПТ» И ОЦЕНКИ ЭФФЕКТИВНОСТИ ЕГО ИСПОЛЬЗОВАНИЯ ДЛЯ ПОИСКА КОЛЛОКАЦИЙ В КОРПУСЕ М. В. ЛОМОНОСОВА[1]

В статье описаны функции и параметры модуля n-грамм информационно-аналитической системы (корпуса) «Манускрипт» и итоги эксперимента по применению нескольких статистических методов в корпусе текстов М. В. Ломоносова. Показано, что количественные и статистические методы оценки биграмм применимы к авторскому историческому корпусу и позволяют выявлять устойчивые сочетания.

Ключевые слова: исторический корпус, корпус Ломоносова, статистические методы, *n*-граммы, меры ассоциации, коллокации.

1. **Цели и задачи**

Вслед за современными корпусами, успешно и продуктивно использующимися для решения самых разнообразных теоретических и практических задач, началось создание исторических и диахронических корпусов русского языка [1–9]. Полнота и возможность оперативного привлечения первичных материалов уже сегодня позволяет ученым решать исследовательские [10, 11] и прикладные задачи [12] традиционными историко-лингвистическими методами.

В то же время большие объемы лингвистических данных созданных исторических корпусов позволяют поставить вопрос о возможности автоматического анализа материала с помощью статистических методов, которые, как известно, успешно применяются сегодня в отношении современных текстов, в том числе и в исследовательских целях.

Целью данной работы является описание возможностей модуля *n*-грамм корпуса произведений М. В. Ломоносова (lomonosov.pro) [13] и демонстрация возможности его использования для выявления устойчивых сочетаний (http://manuscripts.ru/mns/cred_ngr.stat?p_collect=50584966).

Задачи работы:

– описание параметров запросной формы модуля *n*-грамм;

– демонстрация различий при выявлении биграмм тремя статистическими методами;

– предварительная оценка эффективности применения этих трех методов для выявления коллокаций;

– демонстрация различий между биграммами с наиболее высокими значениями в подкорпусах;

– предварительная оценка возможности применения статистических методов для характеристики подкорпусов, различающихся стилистически, жанрово и тематически.

2. **Данные**

Корпус языка Ломоносова (lomonosov.pro) создан на основе полного собрания сочинений в 11 томах (далее – ПСС), изданных в период с 1950 по 1984 г., и открыт 18 ноября 2011 г. Содержит 1150 текстов: 934 текста на русском языке, 101 текст на латинском языке, тексты на немецком, французском, шведском языках. Объем – более 1,1 млн словоформ. Объем базы данных – более 100 Гб.

Для экспериментов были выбраны тексты двух томов – восьмого (объем 96665 словоформ), содержащего художественные произведения, и десятого (объемом 128619 словоформ), в котором размещены письма и автобиографические материалы [14]. Выбор томов обусловлен различными стилем, жанрами и тематикой произведений, включенных в эти томы.

3. **Корпусный менеджер и первая версия модуля n-грамм**

Запросные формы корпуса (http://lomonosov.pro/mns/srch.simple?p_ed_id=50584966) позволяют, построив подкорпус на основе метаданных томов или произведений, получить индексы словоформ, конкордансы, а после лемматизации - указатели начальных форм и полный грамматический указатель, использовать фильтр по маске и/или грамматическим признакам для поиска определенных слов или их форм. Сортированные перечни включают адрес единицы в ПСС и количественную информацию.

В публикации [15] описаны возможности и параметры запроса первой версии модуля *n*-грамм, выходная форма которого позволяла получить сведения об абсолютной частоте встречаемости сочетаний из двух, трех, четырех или более словоформ в одном или одновременно нескольких текстах. Возможность указания масок компонентов сочетаний, расстояния между компонентами, фильтр незнаменательных слов (союзов, предлогов, частиц) и некоторые другие параметры позволяли получить сортированный по

[1] Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (РГНФ), проект № 14-04-00585 «Корпус языка М. В. Ломоносова: аналитическая и лингвистическая разметка».

убыванию количества список сочетаний, часть из которых – частотных, грамматически и/или семантически связанных – может рассматриваться в качестве претендента на статус устойчивого в тексте и/или языке.

Цель дальнейшего совершенствования модуля – предоставить пользователям полноценный и гибкий инструмент анализа данных корпуса с помощью статистических методов.

4. **Параметры запроса модуля *n*-грамм**

В настоящее время запросная форма модуля позволяет:

– сформировать подкорпус текстов;

– указать количество компонентов *n*-граммы;

– указать расстояние между компонентами – как фиксированное (в контакте, через одну словоформу и т. д.), так и нефиксированное (в контакте или на расстоянии одной, двух и т. д. словоформ);

– ввести буквенные образцы или маски компонентов (% – любое количество любых символов, _ – один любой символ);

– использовать при вводе маски регулярные выражения;

– указать единицы, по отношению к которым применяются маски, – словоформа или начальная форма;

– использовать фильтры:

• количественный – вывести на экран сочетания, встретившиеся в подкорпусе указанное количество раз;

• грамматический – указать грамматические значения компонентов (при лемматизации подкорпуса);

• частеречный – не учитывать при построении сочетаний служебные слова – союзы, предлоги, частицы;

– выбрать процедуру анализа:

• подсчет абсолютного количества;

• подсчет относительного количества;

• MI-тест;

• PMI-тест;

• t-score-тест;

• log-likelyhood-тест;

• Dice-тест;

• Chi-squared-тест.

Возможность совмещать значения разных параметров делает запросную форму максимально гибкой для построения запроса.

5. **Предварительная оценка материала и эффективности методов**

5.1. Теоретические вопросы

5.1.1. Авторский исторический корпус

В одной из наших работ сделана попытка обосновать необходимость создания и возможность использования в традиционных историко-лингвистических исследованиях исторических и авторских корпусов, несмотря на то, что их объем и сбалансированность уступают современным, так как эти параметры зависят от количества сохранившихся текстов: «Несмотря на то, что при создании исторического и авторского корпусов часто невозможно достичь такого количества единиц (например, слов или синтаксических конструкций), которое необходимо для получения статистически значимых величин, представленность в таких корпусах всех произведений конкретного автора или всех текстов определенного времени позволяет считать их не менее важными для исследования языка, чем современные корпусы большого объема: авторский корпус репрезентирует подъязык автора, являющийся частью языка конкретной эпохи; исторический корпус в определенном временном диапазоне дает факты для описания языковой системы, сопоставимой с системами предшествующего и последующего периодов.

Таким образом, учет свойств вхождения и сопоставимости позволяют снять ограничения, связанные с недостаточной представленностью в текстах языковых явлений, и даже единичные, статистически непоказательные случаи вариативности рассматривать как достаточно надежные на более широком фоне языка определенного времени или конкретного автора» [16].

Можно предположить, что важные для современного корпуса, но недостижимые для исторических или авторских корпусов параметры – объем и сбалансированность – не будут критичными для последних и при применении статистических методов в том случае, если результаты будут рассматриваться не сами по себе, а в сопоставлении с результатами, полученными на подкорпусах, характеризующимися иными лингвистически, и/или текстологическими, и/или временными характеристиками.

5.1.2. Коллокации

Понятие коллокации как лингвистической единицы обсуждается в большом количестве работ современных лингвистов (из работ последних лет см., например, [17–20]).

Лингвистическое понимание коллокации включает в себя представление о грамматически и/или семантически устойчивом сочетании слов (словоформ). При этом компоненты коллокации (коллокаты) имеют тенденцию к устойчивой (неслучайной) совместной встречаемости (сочетаемость > регулярность > устойчивость). Устойчивость проявляется: а) в ограниченном количестве вторых коллокатов и б) в регулярном совместном использовании. Последнее позволяет использовать для поиска коллокаций количественные и статистические методы (см., например, [21–34]).

5.1.3. *N*-граммы

N-грамма - последовательность из *n* компонентов, извлеченная из текста. Частота появления определенного сочетания двух словоформ (слов) в некотором подкорпусе текстов является одним из критериев устойчивости сочетания. Частота использования может оцениваться как в абсолютных, так и относительных величинах. Вторым критерием устойчивости является регулярность, что выявляется при учете частоты встречаемости каждого из компонентов сочетания, а также при сравнении реального и возможного, вероятного, статистически ожидаемого количества сочетаний в подкорпусе. Оценка устойчивости (речевой и/или языковой, грамматической и/или сти-

листической регулярности) сочетания осуществляется в числовых мерах устойчивости (ассоциации, связности) и вычисляется с помощью статистических методов анализа *n*-грамм.

Весь диапазон вопросов связи *n*-граммы – коллокации – конструкций («последовательность компонентов (= состав) – частота последовательности (= количество) – статистически значимая частота –> сила ассоциации / степень неслучайной совместности (= статистика) – контекстно значимая связь (= целостность; коллокация) – лингвистически значимая связь (конструкции)») рассматривается, например, в [35] и в других работах. В работе также показано, что применение статистических методов решения исследовательских задач должно предваряться формированием подкорпуса на основе стилистических (а также текстологических, жанровых, тематических) характеристик [36].

5.2. Прикладные вопросы

5.2.1. Меры ассоциации

Типология методов, применяемых в лингвистической статистике, представлена в [37]. В этой же работе дана оценка эффективности применения конкретных статистических мер ассоциации для поиска значимых сочетаний. В работах российских исследователей эти оценки подтверждаются и уточняются [38-43].

Меры ассоциации обладают различными возможностями для выделения коллокаций. Так, мера MI (mutual information) «позволяет выделять наиболее редкие и своеобразные коллокации и подходит для выделения терминологии, имен собственных и прочих конструкций, в которых частота составляющих коллокацию слов ничтожно мала» [44]; выявляет «сложные номинации: термины, наименования объектов, ключевые определения предметной области» [45], «чувствительна к случайным совпадениям, находит тематические коллокации» [46].

Мера t-score «оказывается полезна при решении задачи о выделении тех единиц, которые характеризуют *все* (или *подавляющее большинство*) [выделено – авторами] текстов коллекции. Основная масса таких сочетаний характеризует, скорее, особенности стиля текстов коллекции» [47], «позволяет найти наиболее распространенные частые обороты» [48]. «Критерий t-score направлен, прежде всего, на выделение “устойчивых конструкций”, клише и “общеязыковых устойчивых сочетаний” (производных служебных слов, дискурсивных слов)» [49]. В целом метод оценивается как один из лучших для извлечения коллокаций [50].

Особенностью Dice-меры является то, что она «находит симметричные устойчивые сочетания (w_1 и w_2 встречаются только вместе)» [51], что позволяет выявлять в подкорпусе слова с ограниченной сочетаемостью, а соответственно, и сочетания, с высокой степенью вероятности претендующие на статус коллокаций.

5.2.2. Данные

Полнотекстовая база данных авторского исторического корпуса Ломоносова содержит размеченную машиночитаемую копию всех текстов полного собрания сочинений. Тексты (произведения) имеют метаразметку; тексты на русском языке лемматизированы (В настоящее время осуществляется редактирование грамматической базы данных морфологического анализатора для максимально полной лемматизации и корректура текстов для устранения опечаток).

5.3. Методика

В основе эксперимента лежит корпусный подход, основанный на извлечении *n*-грамм без заранее заданных ограничений (лексических, морфологических, синтаксических, семантических, тематических) и предназначенный для выявления частотных и статистически значимых с точки зрения совместной встречаемости словоформ. Для поиска коллокаций этот подход ограничен значительным количеством лингвистически незначимых сочетаний, а соответственно, необходимостью ручной постобработки результатов [52].

Эксперимент проводился над *n*-граммами, состоящими из двух компонентов. Биграммы оценивались с помощью нескольких количественных и статистических мер.

Для эксперимента выбраны следующие меры ассоциации:

– мера MI (коэффициент взаимной информации; из группы точечных оценок силы ассоциации, по [53]), помогающая выявить степень зависимости или независимости компонентов сочетания в подкорпусе,

– мера t-score (коэффициент ассоциации; из группы асимптотических критериев для проверки гипотезы, по [54]), позволяющая определить степень случайности или неслучайности силы связанности (ассоциации) компонентов сочетания;

– мера Dice (коэффициент взаимного ожидания; из группы точечных оценок силы ассоциации, по [55]), дающая высокий показатель для симметричных коллокаций, в которых компоненты встречаются только вместе.

Для сравнения были привлечены данные ранжирования биграмм на основании относительной частоты встречаемости (частости), а также данные перечней биграмм на основе лемм в двух выбранных подкорпусах.

Анализировались и сравнивались первые тридцать позиций списков.

6. **Инструментарий**

6.1. Процедуры

Вычисление значений мер осуществлено автоматически с помощью процедур вычисления соответствующих мер.

$$MI = \log_2 \frac{F(w_1, w_2) \times N}{F(w_2) \times F(w_1)};$$

$$t - score = \frac{F(w_1, w_2) - \frac{F(w_1) \times F(w_2)}{N}}{\sqrt{F(w_1, w_2)}};$$

$$Dice = \frac{2 \times F(w_1, w_2)}{F(w_1) + F(w_2)},$$

где $F(w_1)$ – частота первого коллоката в подкорпусе; $F(w_2)$ – частота второго коллоката в подкорпусе; $F(w_1, w_2)$ – частота коллокации w_1w_2 в подкорпусе; N - общее число словоформ в корпусе [56].

Процедуры разработаны программистом Р. М. Гнутиковым инструментальными средствами Oracle на языке PL/SQL.

6.2. Параметры запроса

После формирования подкорпуса (произведения тома 8 и 10) для получения результата использовались следующие постоянные параметры запроса и их значения:

– тип единицы – *словоформа*;

– маска единиц – % (все);

– количество единиц – *две*;

– расстояние – *0–0*; –

и переменные значения следующих параметров:

– мера – *относительное количество / MI / T-score / Dice*;

- незнаменательные слова - *учитывать / не учитывать предлоги, союзы частицы*;

- количество сочетаний - *показывать / не показывать единичные биграммы*.

7. **Результаты**

7.1. Абсолютная и относительная частота

7.1.1. Ранжирование по частоте встречаемости биграмм в 10-м томе (количество биграмм - 128100) показало, что первые тридцать позиций занимают не только незначимые сочетания некоторых наиболее частотных служебных слов (*и в, и не, и о*, *и с* и под.) и предложно-падежные формы (*для того*, *в канцелярию*, *о том*, *в академии* и др.), но и сложные союзы или их части (*но и*, *затем что*), составное наименование *Академии наук* (2-я позиция), имя собственное *Михайло Ломоносов* (4-я позиция), клише *милостивый государь* (14-я позиция), *вашего сиятельства* (20-я позиция), *вашего превосходительства* (28-я позиция), оформляющие формуляр деловых текстов.

Первые тридцать позиций списка наиболее частых биграмм 8-го тома (количество биграмм - 96378) также занимают сочетания с союзом *и* (*и в, и с, и на* и под.), предложно-падежные сочетания (*к нам, к тебе* и под.), части титульных имен (*императорского величества* – 11-я позиция, *ея императорского* – 15-я позиция, *ея величества* – 18-я позиция), 29-я позицию занимает сочетание *мой дух*, а 7-ю и 30-ю - *о коль* и *о как*.

Можно видеть, что, несмотря на значительную схожесть (и даже их совпадение) первых тридцати биграмм двух томов, существенны и различия, связанные, несомненно, со стилем и жанром произведений: соотношение (т. 8 vs. т. 10) сочетаний с союзом *и* и с предлогами - 21 vs. 12, частость идентичных сочетаний - *и в* 0.00180 vs. 0.00248, *и не* 0.00035 vs. 0.00070, *и с* 0.00074 vs. 0.00048, наличие эмоционально-экспрессивных *о коль, о как* vs. целевых сочетаний *для того, и для, того ради*.

7.1.2. Использование фильтра служебных слов (предлогов, союзов, частиц) увеличивает количество дифференцирующих томы примеров (количество биграмм в т. 8 – 77027, в т. 10 – 102178): если в 8-м томе значительное количество атрибутивных сочетаний *мой дух, всякой час, дух мой, торжественный день, всероссийский престол* и др., то в 10-м – грамматических сочетаний разного типа *а особливо, того ради, между тем, сверх того, то есть, может быть*.

7.1.3. В списках наиболее частых сочетаний лемм (фильтр на предлоги, союзы, частицы) в томах (количество биграмм в т. 8 – 59280, в т. 10 – 64311) противопоставлены (т. 8 vs. т. 10):

– Я ТЫ, ТЫ Я, Я ВЫ, ОН Я, ТЫ МЫ, vs. Ø,

– Я ВИДЕТЬ, ТЫ БЫТЬ, ОН БЫТЬ, БЫТЬ ТЫ, Я БЫТЬ, Я МОЧЬ, ТЫ МОЧЬ, Я ДАТЬ vs. Ø;

– Ø vs. АКАДЕМИЯ НАУК, КАНЦЕЛЯРИЯ АКАДЕМИЯ, КАНЦЕЛЯРИЯ АН;

– Ø vs. ПРАВИТЕЛЬСТВОВАТЬ СЕНАТ, КНИЖНЫЙ ЛАВКА, АКАДЕМИЧЕСКИЙ СОБРАНИЕ, АКАДЕМИЧЕСКИЙ КАНЦЕЛЯРИЯ, РОССИЙСКИЙ ЯЗЫК, ПРОФЕССОРСКИЙ СОБРАНИЕ и др.,

что дифференцирует подкорпусы и лексически, и грамматически, тематически, и стилистически.

7.2. MI-тест

Наиболее высокое значение меры MI получают редкие (единичные) сочетания.

7.2.1. Т. 8 (количество биграмм – 96378; $MI_{1–30}$ = = 16.56071):

– уникальные атрибутивные сочетания: *адским углем, алцейской лирой, алчны мытари, амазонском уборе, ангелы пригожия* и под.,

– имена собственные: *Александроневском монастыре, Анне Иоановне, Апостола Андрея, архиепископ Дюк* и др.;

но и *азийский воскресшу, азийским разъежжая, алчбу претерпевают* и под., которые остаются за пределами устойчивых.

Т. 10 (количество биграмм – 128100; $MI_{1–30}$ = = 16.97274):

– имена собственные: *Александре Петровиче, Алексее Григорьевиче, Алексеем Левским* и др.;

– терминологические сочетания: *анатомические препараты, анатомическом театре*;

– однородные члены: *американских морских, анатомией химией*.

7.2.2. Более показательными являются биграммы, встретившиеся в подкорпусах по два раза и имеющие в т. 8 в пределах первых 30 случаев значение MI от 15.56071 до 14.97574, в т. 10 – 15.97274:

т. 8:

– атрибутивные сочетания: *божием величестве, большем отдалении, витыя ракеты, воплю первому, мягкою травою, пушечной пальбе, римскому обыкновению* и др. (всего – 15);

– имена собственные: *Александру Невскому, Ивану Ивановичу* (и их части – *Ивановичу Шувалову, деля* [Моль]);

– субстантивные сочетания: *ливане кедры, отпущение недоимок, лисицы кожу*;

т. 10:

– имена собственные: *Екатерина Алексеевна, Сарское Село* (их части – *Ларионовичем Воронцовым*);

– сочетания, включающие слова, обозначающие титул, статус, должность и под.: *лаборатора Бетигера, пресвятыя Богородицы, хирург прозектор*;

– атрибутивные сочетания разной степени терминологичности: *анатомического театра, мужеска полу, обыкновенного барометра, пасхальной ярмарки, царская водка*.

7.3. T-score тест

Высокое значение с помощью меры t-score получают *n*-граммы, включающие высокочастотные компоненты. В отличие от ранжирования на основе абсолютного количества встречаемости понижается значения для сочетаний, компонентами которых являются максимально частотные словоформы (слова).

7.3.1. Т. 8 (количество биграмм – 96378; T-score$_{1–30}$ = = 7.30427–4.41335):

– предложно-падежные сочетания со вторым компонентом местоимением: *к нам, к тебе, в том, в нем* и др. (всего 13 случаев);

– предложно-падежные сочетания с существительным: на земли, в след, в свете, на свете;

– сочетания с не: *не может, не токмо, не могут, не мог*;

– титулы: *ея величества* и их компоненты (императорского величества, ея императорскаго);

– сочетания с начальным *о*: *о коль, о как, о боже*.

По сравнению с первыми тридцатью n-граммами, которые были выделены на основе абсолютной частоты встречаемости, нет ни одного сочетания с союзом *и*, большую часть составляют предложно-падежные сочетания, к двум экспрессивно-эмоциональным добавляется *о боже*, а *о коль* с 7-й позиции поднимается на вторую, а о как - с 30-й на 18-ю (см. раздел 7.1).

Т. 10 (количество биграмм – 128100; T-score$_{1–30}$ = 15.19557-6.95871):

– сложные союзы и предложно-падежные формы, часто выполняющие союзную функцию: *но и, затем что, так как, а особливо, для того, между тем, сверх того, о том, в том, к тому*;

– составное наименование: *Академии наук* (1-я позиция);

– имя собственное: *Михайло Ломоносов* (3-я позиция);

– этикетные формулы и титулы и их аббревиатуры: *милостивый государь, вашего сиятельства, вашего превосходительства, е. и., и. в.*;

– не связанные формы: *что он, что я*.

Первые 30 позиций списка занимают устойчивые сочетания грамматического, семантического и клишированного характера, в отличие от ранжированного списка на основании абсолютного количества отсутствуют n-граммы с союзом *и* (см. раздел 7.1).

7.3.2. При удалении из выборок союзов, предлогов частиц повышается ранг номинативных сочетаний.

Т. 8 (количество биграмм – 77029; T-score$_{1–30}$ = = 6.77463–3.11968):

– имена собственные и части титулов: Петра Великаго, Елисаветы Петровны, Великий Петр, императорскаго величества, ея императорскаго, ея величества, государыни императрицы, императрицы Елисаветы;

– атрибутивные сочетания: мой дух, сей день, всякой час, дух мой, торжественный день, российский престол, российский род и др. (всего 17 примеров);

– предикативные сочетания: я вижу, ты можешь;

– наречные сочетания: коль много, коль долго.

В первых 30 сочетаниях увеличивается количество имен собственных и титулов и количество атрибутивных сочетаний, терминологического и обстоятельственного характера.

Т. 10 (количество биграмм – 102178; T-score$_{1–30}$ = = 15.19557–5.37949):

– составные номинации: *Академии наук* (1-я позиция), правительствующего Сената, канцелярии АН, канцелярию АН, канцелярии Академии;

– имена собственные: Михайло Ломоносов (2-я позиция), Иван Иванович;

– титулы, этикетные формулы: милостивый государь, вашего сиятельства и под. (всего 8 случаев);

– союзы, союзные слова, наречные выражения, модальные слова: а особливо, между тем, сверх того, то есть, может быть.

Устранение незнаменательных слов поднимает ранг имен собственных, составных номинаций, грамматически значимых сочетаний.

Показательным с точки зрения оценки связности компонентов в *n*-граммах является числовое значение меры T-score, которая в 10-м томе в несколько раз выше, чем в 8-м.

7.4. Dice-тест

Тест позволяет обнаружить симметричные *n*-граммы, компоненты которых встречаются в подкорпусе только вместе.

7.4.1. Список первых 30 биграмм включает биграммы, которые встретились по одному разу и каждый из компонентов которых в подкорпусе встретился по одному разу.

Т. 8 (количество биграмм – 96378; Dice$_{1–30}$ = = 1.00–0.80):

Список идентичен списку первых 30 биграмм теста MI. Отличие одно: на 6-й позиции Dice появляется Александру Невскому, отсутствующее в MI$_{1–30}$.

Т. 10 (количество биграмм – 128100; Dice$_{1–30}$ = = 1.00)

Список первых 30 биграмм идентичен аналогичному списку теста MI

7.4.2. Первые 30 биграмм, которые встретились по два раза.

Т. 8 (количество биграмм – 96378; $Dice_{1–30} = 1.00$)

Перечень первых 30 биграмм, встретившихся по 2 раза, лишь частично совпадает с аналогичным списком биграмм MI теста:

– в Dice максимальную оценку получают имена собственные: *Ивану Ивановичу* (1.00 vs. 14.97574 при max 15.56071), *Михайло Ломоносов* (0.888889 vs. за пределами первых 30);

– части титулов и этикетных формул: всепресветлейшей державнейшей (1.00 vs. за пределами первых 30-ти), *императорских высочеств* (1.00 vs. vs. 14.97574 при max 15.56071);

– составные номинации: *преложение псалма* (1.00 vs. за пределами первых 30), *перспективном расположении* (0.85714 vs. за пределами первых 30).

Т. 10 (количество биграмм – 128100; $Dice_{1–30} = 1.00$)

Различий между первыми 30 биграмм, встретившимися по 2 раза, в Dice и MI меньше, чем в т. 8, но они того же характера:

– в список первых 30 попадают имена собственные и этикетные и титульные формулы, отсутствующие в списке MI: *Анна Иоановна*, *всепресветлейшая державнейшая*;

– составные номинации: *ботанического сада* (все 1.00 vs. за пределами первых 30 перечня MI).

8. Выводы

Результаты эксперимента позволяют сделать следующие выводы:

– относительно небольшой объем подкорпусов исторического авторского корпуса не является критическим препятствием для получения показательных, соотносящихся с результатами, получаемыми на данных большего объема, результатов и оценок;

– количественные и статистические методы оценки биграмм показывают, что высокие значения мер в большом количестве случаев получают сочетания, компоненты которых семантически и/или грамматически связаны;

– сравнительный анализ результатов применения методов к подкорпусам позволяет увидеть существенные различия в лексической сочетаемости как на уровне состава первых 30 биграмм, так и в отношении частости и значений мер ассоциаций конкретных сочетаний и их видов;

– анализ результатов оценки степени связанности сочетаний статическими методами, применяемыми к историческому или авторскому корпусу, должен сопровождаться ручной постобработкой и осуществляться с учетом:

– объема корпуса;

– наличия шумов, связанных с опечатками;

– неполной лемматизации текстов и некоторых других условий;

– эффективным оказалось применение лингвистических параметров при ранжировании *n*-грамм: перечни наиболее частых *n*-грамм, состоящих из лемм двух подкорпусов, содержащих произведения разных стилей и жанров, оказались различными: эмоциональность и художественное воздействие восьмого тома противопоставлено информационности и событийности текстов десятого тома.

Показательный для дифференциации подкорпусов, содержащих тексты различного стиля и жанра, анализ наиболее частых *n*-грамм, состоящих из лемм, позволяет предположить, что использование лингвистических параметров при формировании рейтинга *n*-грамм (например, учет/неучет порядка следования компонентов, границ синтаксических конструкций, раздельная оценка симметричных и несимметричных, свободных и несвободных *n*-грамм и др.) должно дать показательные результаты, а использование таких параметров – объективный инструмент оценки *n*-грамм.

8. **Перспективы**

Результативность и эффективность применения количественных и статических методов к данным авторского корпуса позволяет наметить дальнейшие шаги по развитию этого направления в рамках проекта «Манускрипт», в частности:

– проверка возможности использования других статистических методов для оценки регулярности и устойчивости *n*-грамм в корпусе Ломоносова;

– оценка эффективности применения количественных и статистических мер ассоциации компонентов *n*-грамм к подкорпусам, сформированным из текстов XI-XV веков;

– добавление структурно-лингвистических параметров формирования выборки и ранжирования *n*-грамм и оценка возможности выявления с их помощью устойчивых сочетаний;

– эксперименты по применению статистических мер к *n*-граммам, состоящим из более чем двух компонентов;

– эксперименты по оценке эффективности использования одновременно статистических методов и структурно-лингвистических параметров и др.

Работы по применению собственно статистических методов в отношении авторских и исторических корпусов в славистике только начинается. И хотелось бы, чтобы эти методы стали точным и надежным инструментом в историко-лингвистических исследованиях.

Библиографические ссылки

1. Национальный корпус русского языка [Электронный ресурс]. – URL: www.ruscorpora.ru (дата обращения: 12.09.2016).

2. *Савчук С. О., Сичинава Д. В., Гарипов И. И.* Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работы [Электронный ресурс]. – URL: http://fccl.ksu.ru/issue_spec/docs/Savchuk_Sichinava_Garipov.doc (дата обращения: 18.09.2016).

3. *Соловьев В. Д., Ахтямов Р. Б.* Корпус русского языка XVIII века: текущее состояние // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам: материалы Междунар. науч. конф., Ижевск, 13–17 июля 2006 г. – Ижевск, 2006. – С. 156-160.

4. *Savchuk, Svetlana.* Corpus-based Investigation of Language Change: the Case of RNC // Proceedings of the Corpus Linguistics Conference CL2007 University of Birmingham, UK, 27-30 July 2007 / Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.). – URL:

http://ucrel.lancs.ac.uk/publications/CL2007/final/181/181_Paper.pdf (дата обращения: 12.09.2015).

5. *Баранов В. А., Аникина Р. А., Кокорина Т. В., Ощепков С. В., Соколова А. А.* Метаинформация в коллекции М. В. Ломоносова на портале «Манускрипт: Славянское письменное наследие» // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: материалы междунар. науч. конф. (Казань, 26–30 августа 2008 г.) / отв. ред. В. А. Баранов, В. Д. Соловьев. – Казань : Изд-во КГУ, 2008. – С 23–27.

6. *Савчук С. О.* Корпус текстов XVIII века в составе Национального корпуса русского языка: проблемы и перспективы // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: материалы Междунар. науч. конф. (Казань, 26-30 августа 2008 г.) / отв. ред. В. Д. Соловьев, В. А. Баранов. – Казань : Изд-во Казан. гос. ун-та, 2008. С. 241–244.

7. *Савчук С. О, Сичинава Д. В.* Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб. : Нестор-История, 2009. – С. 52–70. – URL: http://ruscorpora.ru/sbornik2008/04.pdf (дата обращения: 12.09.2016).

8. *Баранов В. А.* Полное собрание сочинений М. В. Ломоносова в интернете: подготовка электронной коллекции и функциональные возможности модулей корпуса // Уч. зап. Казанского ун-та. Серия: Гуманитарные науки. – Т. 152. – Вып. 6. – 2010. – С. 223–234.

9. *Баранов В. А.* Корпус языка М. В. Ломоносова // Русский язык: функционирование и развитие (к 85-летию со дня рождения заслуженного деятеля науки Российской Федерации профессора Виталия Михайловича Маркова) : материалы Междунар. науч. конф. (Казань, 18-21 апреля 2012 г.) / Казан. ун-т; Ин-т филологии и искусств; Каф. ист. рус. яз. и слав. языкозн. ; под общ. ред. Л. Р. Абдулхаковой, Д. Р. Копосова. – Казань : Казан. ун-т, 2012. – Т. 1. – С. 58–63.

10. *Баранов В. А.* Историческая морфология и корпусная лингвистика: стяженные и нестяженные формы имен в русских рукописях XI века // Русский язык: история и современность: сб. ст. к юбилею проф. Т. М. Николаевой / под общ. ред. Л. Р. Абдулхаковой, Д. Р. Копосова. – Казань : Казан. гос. ун-т, 2008. – С. 43–53.

11. *Сичинава Д. В.* Исторические корпуса Национального корпуса русского языка как инструмент диахронических исследований грамматики // Писменото наследство и информационните технологии [Текст]: материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отв. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. – София ; Ижевск, 2014. – С. 226–229.

12. *Баранов В. А., Гнутиков Р. М., Зливко С. Д.* Авторский электронный словарь-справочник лингвистической терминологии М. В. Ломоносова // Интеллектуальные системы в производстве. – 2015. – № 3 (27). – С. 88–92.

13. Корпус М. В. Ломоносова [Электронный ресурс] / Ижевский государственный технический университет, кафедра лингвистики, Центр теоретической и прикладной лингвистики, 2005–2016; Казанский (Приволжский) федеральный университет, 2007–2009; Удмуртский государственный университет, лаборатория по автоматизации филологических работ, 1989–2013; рук. проекта В. М. Марков, 1989–2010; соурук. и рук. В. А. Баранов, 1989–2016. – URL: lomonosov.pro (дата обращения: 18.09.2016).

14. *Ломоносов М. В.* Полное собрание сочинений : в 11 т. – Т. 8: Поэзия. Ораторская проза. Надписи. 1732–1764 гг. М.; Л., 1959. 1280 с.; Т. 10: Служебные документы. Письма. – М. ; Л., 1959. – 935 с.

15. *Баранов В. А.* Организация поиска и демонстрации коллокаций в корпусе «Манускрипт» // Проблемы истории, филологии, культуры. – 2014. – № 3 (45). – С. 275–277.

16. *Баранов В. А.* Полное собрание сочинений М. В. Ломоносова в интернете... – С. 225.

17. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М. : МИЭМ, 2011. – 272 с.

18. *Влавацкая М. В.* Понятия коллокации и коллигации в диахроническом рассмотрении // Актуальные проблемы филологии и методики преподавания иностранных языков. – 2011. – № 5. – С. 19–25.

19. *Пивоварова Л. М., Ягунова Е. В.* От коллокаций к конструкциям // Acta Linguistica Petropolitana. Труды института лингвистических исследований. – 2014. – Т. 10. – № 2. – С. 568–617. – URL: http://elibrary.ru/download/84557015.pdf (дата обращения: 12.09.2016).

20. *Влавацкая М. В.* Комбинаторная лексикология: функционально-семантическая классификация коллокаций // Филологические науки. Вопросы теории и практики. – 2015. – № 11–1. – С. 56–60. – URL: http://elibrary.ru/download/54324014.pdf (дата обращения: 12.09.2016).

21. *Evert S.* Association Measures [Электронный ресурс] // Computational Approaches to Collocations. – URL: http://collocations.de/AM/index.html (дата обращения: 12.09.2015).

22. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы / под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. – Хельсинки, 2008. – С. 343–357. – URL: https://drive.google.com/file/d/0BwBejXXryRcRSFF2ekRs Y3VYV00/view?pref=2&pli=1 (дата обращения: 12.09.2016).

23. *Хохлова М. В.* К вопросу изучения сочетаемости и устойчивости лексических единиц автоматическими методами // Структурная и прикладная лингвистика. – 2010. – № 8. – С. 206–218.

24. *Хохлова М. В.* Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов). Автореф. дис. ... канд. филол. наук. 10.02.21. – СПб., 2010. – 27 с. – URL: http://dlib.rsl.ru/viewer/01004855815#previewTab?page=1 (дата обращения: 12.09.2016).

25. *Ягунова Е. В., Пивоварова Л. М.* Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Сб. НТИ. – Сер. 2, № 6. – М., 2010. – URL: http://medialing.spbu.ru/upload/files/file_1394529742_4311.pdf; https://goo.gl/PYqRJp (дата обращения: 12.09.2016).

26. Автоматическая обработка текстов…

27. *Захаров В. П., Хохлова М. В.* Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов // Структурная и прикладная лингвистика. – 2012. – № 9. – С. 222-233. – URL: http://elibrary.ru/download/ 35845010.pdf (дата обращения: 12.09.2016).

28. *Ягунова Е. В., Пивоварова Л. М.* От коллокаций к конструкциям // Русский язык: конструкционные и лексико-семантические подходы / Отв. ред. С. С. Сай. – СПб., 2013. – (Acta Linguistica petropolitana. Труды Института лингвистических исследований РАН / Отв. ред. Н. Н. Казанский, Е. В. Ягунова, Л. М. Пивоварова.) – URL: https://goo.gl/tIHeoR (дата обращения: 12.09.2016).

29. *Залесская В. В.* Программа выявления в тексте двучленных статистически значимых осмысленных коллока-

ций (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-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Библиотека Российской Академии наук. 2014. – С. 283-289. – URL: http://ojs.ifmo.ru/index.php/IMS/article/viewFile/267/263 (дата обращения: 12.09.2016).

30. *Захаров В. П., Хохлова М. В.* 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-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Библиотека Российской Академии наук. – 2014. – С. 290-293. – URL: http://ojs.ifmo.ru/index.php/IMS/article/viewFile/268/264 (дата обращения: 12.09.2016).

31. *Кочеткова Н. А.* Статистические языковые методы. Коллокации и коллигации [Электронный ресурс] // Cyberleninka.ru. – URL: http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii (дата обращения: 12.09.2016).

32. *Пивоварова Л. М., Ягунова Е. В.* От коллокаций к конструкциям...

33. *Бобкова Т.* Извлечение коллокаций из корпуса украинских текстов // Computational linguistics / Kompiuterinė lingvistika. – № 27. 2015. – 93–105. – URL: http://www.vpa.ktu.lt/index.php/KStud/article/viewFile/13747/7329 (дата обращения: 12.09.2016).

34. *Хохлова М. В.* Большие корпусы и частотные существительные: предварительные наблюдения // Структурная и прикладная лингвистика. – 2015. – № 11. – С. 174–185.

35. Автоматическая обработка текстов... – С. 23–25.

36. Там же. – С. 43.

37. *Evert S.* Association Measures.

38. Автоматическая обработка текста...

39. *Захаров В. П., Хохлова М. В.* Выделение терминологических словосочетаний...

40. *Пивоварова Л. М., Ягунова Е. В.* От коллокаций к конструкциям...

41. *Ягунова Е. В., Пивоварова Л. М.* Природа коллокаций в русском языке...

42. *Хохлова М. В.* Исследование лексико-синтаксической сочетаемости...

43. *Хохлова М. В.* Большие корпусы и частотные существительные...

44. *Кочеткова Н. А.* Статистические языковые методы... – С. 302.

45. *Ягунова Е. В., Пивоварова Л. М.* От коллокаций к конструкциям...

46. *Пентус М., Пиперски А., Сорокин А.* Математические модели в лингвистике. Коллокации и их автоматическое определение: лекции. – URL: https://goo.gl/NALvX4 (дата обращения: 12.09.2016).

47. *Ягунова Е. В., Пивоварова Л. М.* От коллокаций к конструкциям...

48. *Кочеткова Н. А.* Статистические языковые методы... – С. 302.

49. *Пивоварова Л. М., Ягунова Е. В.* От коллокаций к конструкциям… – С. 569.

50. *Evert S.* Association Measures. Section 4.3.

51. *Пентус М., Пиперски А., Сорокин А.* Математические модели в лингвистике. Коллокации и их автоматическое определение...

52. *Бобкова Т.* Извлечение коллокаций из корпуса украинских текстов... – С. 98.

53. *Evert S.* Association Measures.

54. Там же.

55. Там же.

56. Там же.

* * *

V. A. Baranov, Doctor of Philology, Professor, Kalashnikov ISTU

Experience of Creation of the N-Gram Module of the System “Manuscript” and Evaluation of The Efficiency of Its Application to Search Collocations in the Corpus of M.V. Lomonosov

The article contains a description of functions and parameters of the n-gram module of the informational analytical system (corpus) “Manuscript” and the results of the experiment on the application of some statistic methods to the corpus of texts by M. V. Lomonosov. The quantitative and statistic methods of evaluation of bigrams are shown as applicable to the author’s historical corpus and enabling revelation of stable combinations.

Keywords: historical corpus, corpus of Lomonosov, statistic methods, n-grams, measure of association, collocations.

Получено: 29.09.16