Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2023»

14–16 июня 2023 г.

Frequency dynamics as a criterion for differentiating inflection and word formation (in relation to Russian aspectual pairs)

Elena V. Gorbova
independent researcher
elenagorbova12@gmail.com

Oksana Iu. Chuikova
Herzen State Pedagogical
University of Russia
oxana.chuykova@gmail.com

Abstract

The paper reports the results of the critical evaluation of the quantitative approach to the distinction between inflection and word formation through the analysis of the trends in the frequency of word forms. The possibility of such analysis is provided by voluminous corpus data and tools for visualizing these trends. Both theoretical foundations of the proposed approach and the results of the pilot study of its applying to Russian aspectual triplets were considered. These cast doubt on the validity of distinguishing between inflection and word formation based on the trends in the frequency of word forms as a reliable tool used to reveal the unity or difference of lexical semantics and thus to define textual units as belonging to the same or different language units.

Keywords: Russian aspect; inflection; word formation, quantitative analysis; frequency, corpora
DOI: 10.28995/2075-7182-2023-22-142-160

Динамика частотности как критерий разграничения словоизменения и словообразования (применительно к видовой парности русского глагола)

Горбова Елена Викторовна
независимый исследователь
elenagorbova12@gmail.com

Чуйкова Оксана Юрьевна
РГПУ им. А. И. Герцена
oxana.chuykova@gmail.com

Аннотация

В статье представлены результаты критического осмысления количественного подхода к проведению границы между словоизменением и словообразованием через анализ динамики частотности употребления словоформ, возможность которого обеспечивается объемными корпусными данными и средствами их визуализации. Предлагается обсуждение как теоретических основ предложенного подхода, так и результатов пилотного исследования его применения к материалу русских видовых троек. Проведенный анализ позволяет усомниться в валидности разграничения словоизменения и словообразования через динамику частотности как действенного способа установления единства лексической семантики в качестве инструмента определения статуса текстовых единиц как вариантов одной языковой единицы или репрезентантов разных.

Ключевые слова: вид русского глагола; словоизменение; словообразование; количественный анализ; частотность, корпус

1 Вступительные замечания

Целью статьи является критическое обсуждение предложенного в [15, 16] «частотного подхода к лексической семантике», или «объективного численного подхода» [15: 73], как применительно к более общему вопросу различения словоизменения и словообразования, так и к его приложению к двум морфологическим типам видовой пары и проблеме трактовки грамматической категории

русского вида (аспекта)[1]. Отметим, что обозначенный подход основан, как указано в [16], на дистрибутивной гипотезе, изложенной в [13].

Далее работа выстроена следующим образом: раздел 2 представит частотный подход к лексической семантике и его критику с точки зрения теоретической лингвистики; в разделе 3 приведены случаи наличия несомненного (по общепризнанным лингвистическим критериям) статуса двух (или более) словоформ одной лексемы при значительно разнящейся частоте употребления словоформ в парадигме одной лексемы и случаи синхронного изменения частотности для словоформ разных лексем; в разделе 4 обсуждаются результаты частотного подхода к лексической семантике на материале видовых троек; в заключительном разделе 5 подведены итоги.

2 Частотный подход к лексической семантике: теоретические основы и его приложение к описанию категории вида русского глагола

Кратко изложим предложенный в [15, 16] подход. Основной задачей [16] является решение «исследовательского вопроса: в какой степени меняется семантика слов при суффиксальном и префиксальном способах образования аспектуальных пар?» [16: 1117]. При этом «проверяемая гипотеза» сформулирована так: «семантическая близость между глаголами в аспектуальных парах перфектив – вторичный имперфектив будет больше, чем между глаголами в парах базовый имперфектив – перфектив» [Ibid]. Постановка задачи и гипотеза базируются на: а) стремлении подвергнуть критическому анализу одну из теоретических моделей русского вида, а именно трактовку первого типа пар (с участием префигированных перфективов, противопоставленных симплексам-имперфективам) как словообразования и второго типа пар (единожды префигированного перфектива и образованного от него вторичного имперфектива) как словоизменения [16: 1116]; б) отказа от решения проблемы словоизменение vs. словообразование на основе обязательности и регулярности, поскольку регулярность отличается градуальностью (авторы ссылаются, в частности, на [11]), см. [16: 1116-1117]. В результате авторы [16] предлагают решать вопрос о границе между словоизменением и словообразованием[2] исключительно на основе идеи о единстве лексической семантики для словоформ одной лексемы, см. [16: 1118-1120], причем последнее устанавливается в ходе анализа изменения во времени частотности употребления словоформ, выявленного путем обращения к корпусу Google Books Ngram (GBN) с визуализацией посредством сервиса Ngram Viewer (https://books.google.com/ngrams/): частота употребления двух лексов с одинаковой лексической семантикой, находящихся в словоизменительном отношении, должна изменяться синхронно [16: 1120].

В [15] акценты чуть сдвинуты: здесь в фокусе внимания и критики находится операционный критерий установления видовой парности, известный как «критерий Маслова». Вместо него формулируется другой критерий видовой парности: «Глаголы из пары «первичный имперфектив – перфектив» имеют одинаковую лексическую семантику (т. е. образуют аспектуальную пару) тогда и только тогда, когда частотность их использования меняется синхронно – графики частотности имеют одинаковую форму» [15: 76]. Кроме визуального сравнения формы кривых на графике, построенном сервисом Ngram Viewer, применяется также их оценка через коэффициент Спирмена. Любопытно, что в [Ibid] авторы приводят три ограничения предложенного метода, и первым идет признание его приблизительности: «критерий не всегда дает точный ответ, т. е. он является не строгим, а приближенным» [Ibid], или градуальности. То есть то, на основании чего в [15, 16] отвергается общепринятый в современной морфологии подход к проблеме словоизменение vs. словообразование через регулярность и обязательность, имея в виду градуальность регулярности.

Заканчивая обзор предложенного в [15, 16] подхода к теоретическому описанию русского вида и критериям видовой парности, остановимся на результатах их исследования. С одной стороны, в [16] авторы приходят к выводу: «основной признак отнесения грамматической категории к

[1] Вполне солидаризируясь с позицией автора [2] относительно сомнительной валидности для лингвистики данных онлайн-инструментов автоматической обработки цифровых текстов, в том числе Google Books Ngram Viewer, мы будем вынуждены, вслед за авторами [15, 16] повторить их исследовательские приемы, чтобы выявить уязвимость предложенного способа решения ряда проблем теоретической лингвистики.

[2] Словоизменение и словообразование и в [15], [16], и в данной работе, понимаются вполне традиционно для отечественной лингвистики, как, напр., в [9], [12], [17], [7].

словоизменению или словообразованию – сохранение или не сохранение лексической семантики – не позволяет разграничить префиксальный и суффиксальный способы видообразования с точки зрения их грамматического статуса. Многолетнюю дискуссию о противопоставлении грамматического статуса аспектуальных пар, образованных префиксальным и суффиксальным способами, по нашему мнению, можно считать практически завершенной» [Ibid: 1127-1128]. С другой стороны, изложенные в [15: 78] результаты анализа графиков изменения частот употребления словоформ (инфинитивов) в парах «базовый глагол – естественный перфектив» и «базовый глагол – специализированный перфектив» (далее – приложение 1 и приложение 2), построенных сервисом Ngram Viewer, на наш взгляд, не дают оснований для таких выводов. Авторы отмечают, что «в приложении 1 из 101 пары в 85 случаях (85%) имеет место высокая (согласно шкале Чеддока) $r > 0,7$ корреляция. Среднее значение коэффициента корреляции 0,780. В приложении 2 из 101 пары в 57 случаях (57%) имеет место высокая $r > 0,7$ корреляция. Среднее значение коэффициента корреляции 0,607» [Ibid]. Ниже авторы отмечают: «вероятность получения ложноположительного решения (неаспектуальная пара показывает высокий коэффициент корреляции), существенно выше, чем ложноотрицательного (аспектуальная пара имеет низкий коэффициент корреляции)» [Ibid]. Однако это можно проинтерпретировать и иным образом: более чем в половине случаев выборки (57%) с парами типа «базовый глагол – специализированный перфектив» результат аналогичен подавляющему большинству (85%) случаев в выборке «базовый глагол – естественный перфектив». Это означает, что большинство случаев в каждом типе пар, постулируемых различными, однако идентичных по морфологической структуре (СВ2 ← НСВ1), отличаются высокой степенью семантической общности, а меньшая их часть (43% и 15% соответственно) показывают существенные семантические различия, отражаемые, по мысли авторов, в несинхронном изменении частотности. С нашей точки зрения, сомнительно, что такой результат исследования дает основания для приведенного выше вывода.

Обратимся к «дистрибутивной гипотезе» как основе предложенного «частотного подхода». В [13], той публикации, на которую ссылаются авторы обсуждаемого подхода как на теоретический источник, находим уточненную формулировку гипотезы: «A distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words» [Ibid: 40]. Пафос [13] заключается как раз в том, чтобы показать, что дистрибутивная модель выстроена на теоретической основе структурализма: множественное цитирование З. Харриса, упоминание Л. Блумфильда, обращение к обоим как к последователям основоположника структурализма Ф. де Соссюра с его понятием значимости языковой единицы и двух видов отношений – синтагматических и парадигматических, что отражено и в приведенной формулировке гипотезы. Речь, следовательно, идет именно о той теоретической основе, которая дает нам используемый, в частности, в отечественной лингвистике подход к различению алло-единиц и «эмических единиц» (вариантов одной сущности и разных сущностей) через анализ дистрибуции и ее типов: дополнительной (непересекающейся) и свободного варьирования, с одной стороны, и контрастной, с другой, см. [9: 198-201]. Дистрибутивный анализ в конечном счете лежит в основе и решения проблемы словоизменение vs. словообразование через регулярность и обязательность, характерные для словоизменения и нехарактерные для словообразования (см. [11: 249-250, 283; 17: 25-26; 13: 51; 10: 326-332; 9: 198-201], т. е. присутствует в том подходе, который отвергается в [15, 16] как недостаточно эффективный из-за градуальности свойства регулярности (кстати, в [12] обязательность также градуальна). Итак, сомнительны как противопоставление дистрибутивной гипотезы в качестве альтернативы решению проблемы словоизменение vs. словообразование через свойства обязательности и регулярности, так и выводимая прямо из этой гипотезы (с отсылкой к [13]) идея об установлении единства лексической семантики на основе синхронности диахронического изменения частот употребления словоформ одной лексемы в качестве единственного критерия словоизменения, см. [16: 1118-1120].

3 Всегда ли словоформы одной лексемы демонстрируют синхронную частотность?

Остановимся на идее авторов [15, 16], согласно которой словоформы одной лексемы демонстрируют синхронную частотность своих вхождений, в то время как частоты вхождений различных

лексем, связанных словообразовательно, изменяются не синхронно. Данная идея подается авторами как сама собой разумеющаяся, без серьезной эмпирической проверки. Думается, что это утверждение представляет собой гипотезу, требующую доказательств. Наблюдаемая частотность вхождений словоформ в парадигмах словоизменительных категорий ставит высказанную идею под сомнение[3]. Приведем примеры.

NB! В первую очередь, следует отметить, что вводимое авторами [15, 16] ограничение по времени (в одном случае установлен временной диапазон с 1920 по 2019 гг., в другом – с 1950 по 2019 гг.) представляется недостаточно оправданным. С одной стороны, на аргумент, согласно которому это делается во избежание влияния старой орфографии, можно возразить, что если изменения в орфографии касаются рассматриваемых глаголов, то, как правило, они затрагивают оба, то есть существенно повлиять на синхронную частотность двух глаголов данный фактор вряд ли может (при этом в нескольких примерах ниже, где наблюдается резкое изменение тренда для определенной словоформы около 1920-го года, рассматривается суммарная частотность для двух вариантов написания). В то же время сокращение (до 100 или 70 лет) временного промежутка, на котором рассматривается динамика частотности глаголов, не позволяет выявить никаких значимых изменений в случае грамматических категорий, изменяющихся медленно. Поскольку было эмпирически установлено, что в русских текстах первой трети XIX в. наблюдаются большие и труднообъяснимые выбросы для любой словоформы, нами было принято решение рассматривать динамику частотности начиная с 1830 г.

1) Вряд ли вызывает сомнение принадлежность к одной словоизменительной парадигме глагольных форм, различающихся временем и лицом. Так, например, словоформы *был*, *буду* и *будет* относятся к одной парадигме, но их частотность в GBN меняется не синхронно, см. рис. 1.

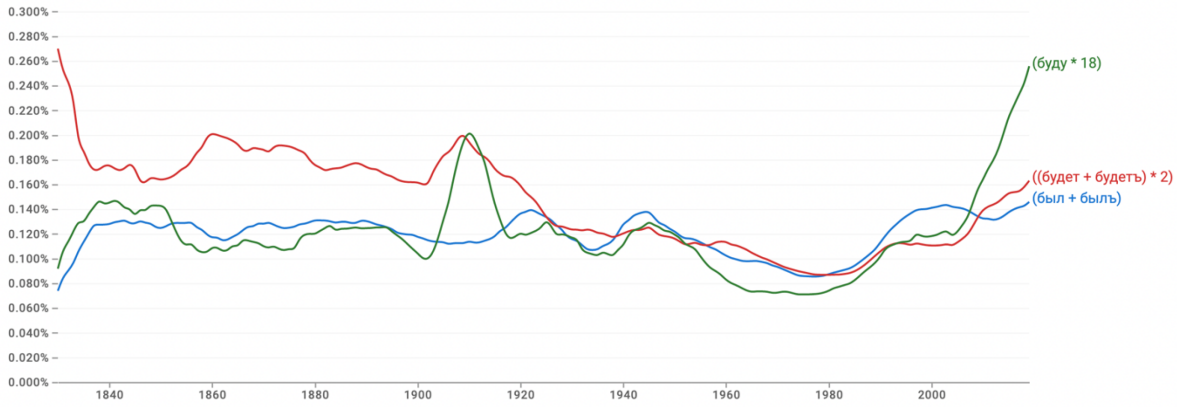

Рисунок 1: Графики частот *был – буду – будет*[4,5]

Для форм *будет* и *буду* на временном промежутке с 1830 по 2019 г. коэффициент корреляции будет невысоким (коэффициент Спирмена, r = 0,47; коэффициент Пирсона, r = 0,42)[6].

Отметим также, что утверждение о синхронном изменении частотности форм одного слова прямо противоречит тому, что известно о развитии граммем футурума и подтверждается на корпусном материале. Выражения со значением намерения, приобретая способность указывать на предсказание (но не наоборот), имеют тенденцию развиваться в граммемы с общим значением футурума (сохраняя способность к выражению значения намерения) см. [6: 310; 5: 106]. В [3] различные пути грамматикализации футурума имеют общую часть, завершаясь семантическим переходом INTENTION > FUTURE. Граммема футурума проходит стадию значения намерения — сначала говорящего, затем агенса высказывания. Далее намерение становится частью значения футурума, а значение предсказания развивается в результате переосмысления намерения третьего лица со стороны говорящего. Распределение форм 1-го и 3-го лица позволяет судить о

[3] Отметим, что статистической единицей в GBN (и в Ngram Viewer) является отдельная словоформа в ее конкретном графическом облике (при невозможности снятия омонимии и различения значений при полисемии).
[4] Степень сглаживания (“smoothong”) задается по умолчанию.
[5] Как и в [15], в случае различий в общей частотности словоформ, для наглядности и удобства сопоставления на графике частоты выравниваются путем кратного увеличения показателей низкочастотных единиц.
[6] Пример расчета коэффициентов корреляции приведен в Приложении Б.

способности граммемы к выражению значений намерения и предсказания и на этом основании судить об этапе развития, на котором находится граммема футурума, см. [6, 4].

На рис. 2 приведен аналогичный случай существенных различий в частотности употребления презентной и претеритальной формы английской проспективной конструкции *be going to+ Inf.*

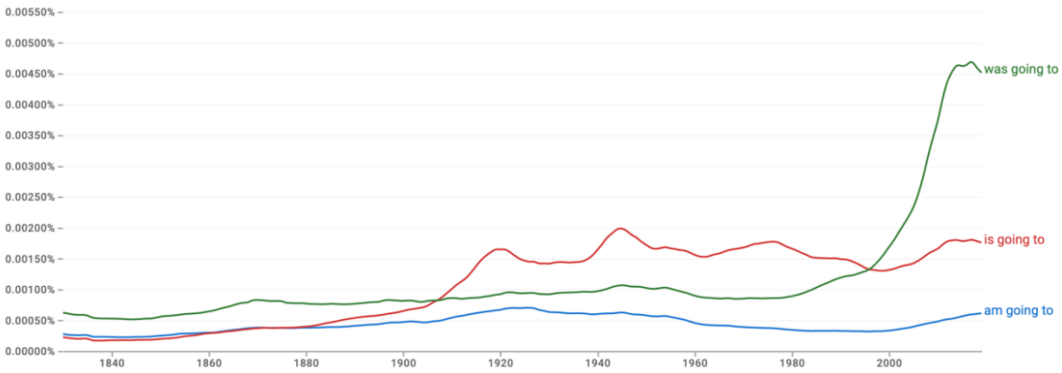

Рисунок 2: Графики частот *am – is – was going to+ Inf.*

Для форм *am going to* и *was going to* на временном промежутке 1830–2019 г. коэффициент корреляции средний или низкий (коэффициент Спирмена, r = 0,53; коэффициент Пирсона, r = 0,27).

2) С другой стороны, есть случаи, в которых высокий коэффициент корреляции наблюдается для глагольных единиц, которые не только не являются членами видовой пары (потенциально, словоформами одной лексемы), но и не связаны деривацией. Примерами служат ингестивные глаголы СВ *съесть* и *выпить* и делимитативы *поесть* и *попить,* относящиеся к одной семантической группе, но обозначающие различные ситуации и, несомненно, являющиеся разными лексемами.

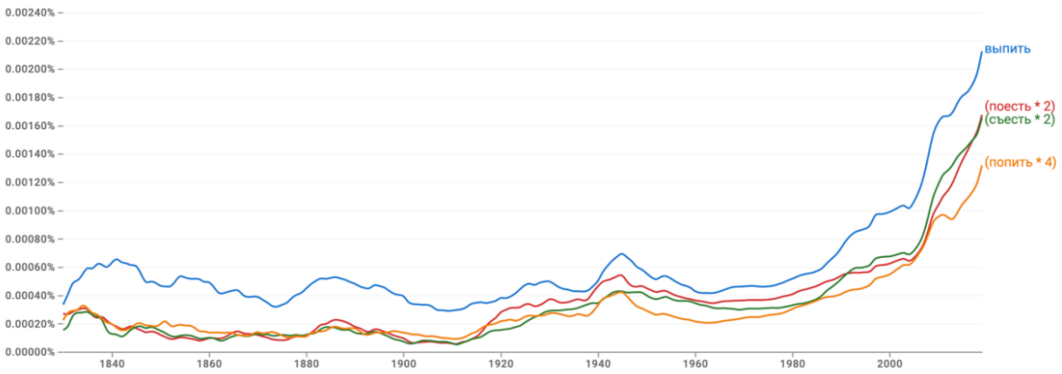

Рисунок 3: Графики частот *съесть – выпить – поесть – попить*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
выпить / попить	0.79	0.97
выпить / съесть	0.75	0.96
выпить / поесть	0.75	0.93
попить / съесть	0.93	0.98
попить / поесть	0.94	0.97
поесть / съесть	0.98	0.99

Таблица 1: Попарные коэффициенты корреляции для *съесть – выпить – поесть – попить*

Следует отметить, что довольно высокий коэффициент корреляции наблюдается для всех (!) парных комбинаций из четырех глаголов, в том числе для *попить*/*съесть* и *выпить*/*поесть*, не обнаруживающих единства ни корня (носителя лексического значения), ни префикса.

На рис. 4 и 5 приведены аналогичные случаи не из области глагольной морфологии: на рис. 4 на материале личных местоимений, на рис. 5 – на однокоренных существительных. В случае местоимений словоформы демонстрируют низкую корреляцию по частотности, а в случае разных, хотя и однокоренных, существительных – высокую (в терминах [15] – ложноотрицательный и ложноположительный результаты).

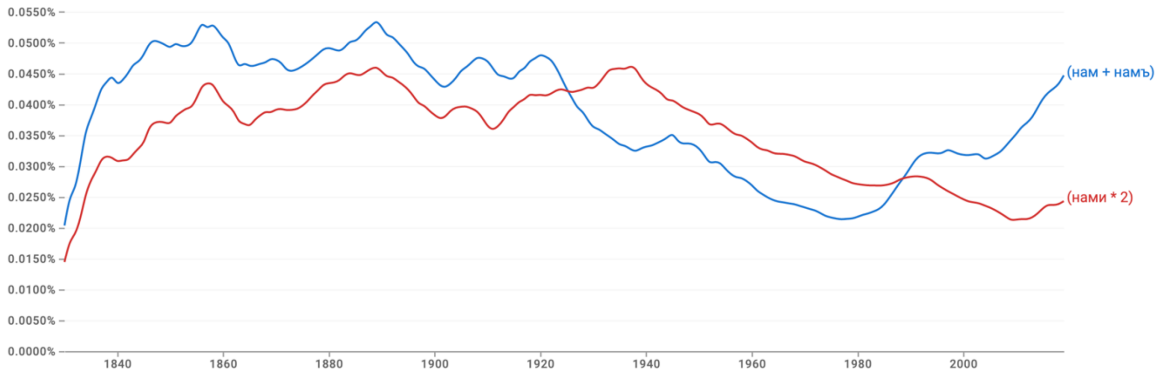

Рисунок 4: Графики частот местоименных форм *нам – нами*

Для форм *нам* и *нами* на промежутке 1830–2019 коэффициент корреляции невысокий (коэффициент Спирмена, r = 0,6; коэффициент Пирсона, r = 0,58).

Рисунок 5: Графики частот существительных *кот – кошка*

Для словоформ *кот*[7] и *кошка* на промежутке 1830–2019 коэффициент корреляции высокий (коэффициент Спирмена, r = 0,82; коэффициент Пирсона, r = 0,9).

3) Авторы делают оговорку о возможности как ложноположительного («неаспектуальная пара показывает высокий коэффициент корреляции»), так и ложноотрицательного результата (аспектуальная пара имеет низкий коэффициент корреляции») [15: 78], при этом вероятность получения ложноположительного результата оказалась выше, чем ложноотрицательного. Ценность такого вывода была бы значительно выше при понимании того, каким образом ложноположительные и ложноотрицательные результаты распределены за пределами рассматриваемых пар. Наличие ложных результатов обоих типов подтверждается приведенными выше примерами. Однако отсутствие данных о системе не позволяет ответить на вопросы: 1) действительно ли распределение ложноположительных и ложноотрицательных результатов дает основание принимать гипотезу о синхронном изменении частотности при словоизменении; 2) можно ли говорить о том, что распределение ложноотрицательных и ложноположительных результатов в [15, 16] соотносимо с таковым в системе языка в целом.

[7] В [2] подмечена неожиданная картина частотности графического слова *кот* в текстах до 1920-х гг., в связи с чем его корреляция со словоформой *кошка* тем более необъяснима.

4 Дистрибутивная гипотеза и частотный подход к видовым тройкам

Еще один ракурс валидности обсуждаемого подхода к решению проблемы словоизменение vs. словообразование через оценку синхронности изменений частотности словоформ в диахронии на больших данных можно увидеть, применив предложенный подход к видовым, или морфологическим биимперфективным, тройкам типа СВ – НСВ1/НСВ2: *съесть – есть/съедать* [18: 235], в которых оба НСВ «претендуют на роль видового коррелята» к СВ [Ibid: 236][8,9]. Будем ориентироваться на именно так понимаемые видовые тройки, причем исключительно «образцовые» [Ibid: 241]: те, в которых оба НСВ, бесприставочный и приставочный (каждый из них реален, а не потенциален), претендуют на роль коррелята к приставочному СВ[10]. При этом все три единицы имеют общий корень и способны обозначать одну и ту же ситуацию, в силу чего характеризуются единством лексического значения.

В [8] представлено диахроническое исследование троек на материале НКРЯ (https://ruscorpora.ru/). Поскольку есть возможность опереться на результаты упомянутого исследования, из десяти троек в [8] отберем шесть – те, для которых частотность НСВ2 не исчезающе низка: *свариться — вариться / свариваться*[11] (также без *-ся: сварить — варить / сваривать*), *съесть — есть / съедать, оторвать — рвать / отрывать, пробить — бить / пробивать, сгореть — гореть / сгорать, сорвать — рвать / срывать*, добавив к ним тройки *разбить — бить / разбивать*, *разорвать — рвать / разрывать*, *намазать — мазать / намазывать, налить — лить / наливать*. Тем самым материалом настоящего исследования являются 11 видовых троек.

Рассмотрим коэффициенты корреляции для всех теоретически возможных парных комбинаций в тройках: СВ2/НСВ1, СВ2/НСВ2 и НСВ1/НСВ2. В табл. 2 и 3 приведены коэффициенты корреляции Спирмена для трех парных комбинаций (без учета порядка следования) в 11 рассматриваемых тройках. Соответствующие графики и расчеты коэффициентов корреляции Спирмена и Пирсона приведены в Приложении А.

Аспектуальная тройка	**СВ2 / НСВ1**	**СВ2 / НСВ2**	**НСВ1 / НСВ2**
съесть — есть / съедать	0.87	0.88	0.77
налить — лить / наливать	0.15	0.88	0.31
разорвать — рвать / разрывать	-0.28	0.88	-0.53
сорвать — рвать / срывать	0.69	0.83	0.84
намазать — мазать / намазывать	0.73	0.8	0.73
оторвать — рвать / отрывать	0.82	0.71	0.41
пробить — бить / пробивать	0.26	0.71	0.27
сгореть — гореть / сгорать	0.89	0.67	0.74
сварить — варить / сваривать	0.89	0.58	0.44
разбить — бить / разбивать	0.13	0.42	0.42
свариться — вариться / свариваться	0.38	0.16	-0.03
Среднее	**0.50**	**0.68**	**0.40**
Медиана	**0.69**	**0.71**	**0.42**

Таблица 2: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ2)

Коэффициент корреляции < 0.7 в парах СВ2 / НСВ2 наблюдается для троек, где НСВ2 низкочастотен (*сваривать(ся)*) и/или НСВ1 и НСВ2 проблемно взаимозаменяемы (*гореть / сгорать*), или же коэффициент низкий, но выше, чем для СВ2 / НСВ1 (*разбить — бить / разбивать*).

[8] Общая оценка видовых троек: это «неотъемлемая составляющая русской аспектуальной системы. <…> Тот факт, что многие русские приставочные глаголы сов. вида вступают в <…> корреляцию с двумя разными глаголами несов. вида, означает только то, что в этом участке системы, помимо собственно аспектуальной корреляции, в игре участвуют также парадигматические отношения лексической синонимии» [18: 247].

[9] В [14], более ранней работе, выполненной тем же исследовательским коллективом, что и в [15, 16], и с применением того же Ngram Viewer, рассматриваются видовые тройки (триплеты) с точки зрения их эволюции за два века. С выводом авторов о том, что «доля глаголов несовершенного вида уменьшается, а вторичные имперфективы вообще вымываются из языка» [14: 425], данные нашего исследования согласиться не позволяют.

[10] Вслед за [7] обозначим приставочный СВ как СВ2, считая первичным СВ, т. е. СВ1, перфектив-симплекс типа *дать*.

[11] Графическая подача и порядок следования членов тройки соответствует принятому в [18: 242-247].

Аспектуальная тройка	СВ2 / НСВ1	СВ2 / НСВ2	НСВ1 / НСВ2
сварить — варить / сваривать	0.89	0.58	0.44
сгореть — гореть / сгорать	0.89	0.67	0.74
съесть — есть / съедать	0.87	0.88	0.77
оторвать — рвать / отрывать	0.82	0.71	0.41
намазать — мазать / намазывать	0.73	0.8	0.73
сорвать — рвать / срывать	0.69	0.83	0.84
свариться — вариться / свариваться	0.38	0.16	-0.03
пробить — бить / пробивать	0.26	0.71	0.27
налить — лить / наливать?	0.15	0.88	0.31
разбить — бить / разбивать	0.13	0.42	0.42
разорвать — рвать / разрывать	-0.28	0.88	-0.53
Среднее	**0.50**	**0.68**	**0.40**
Медиана	**0.69**	**0.71**	**0.42**

Таблица 3: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ1)

Общие наблюдения:

1) по медианному уровню коэффициента корреляции Спирмена из трех возможных пар глаголов, формируемых на базе видовой тройки, наиболее высок уровень корреляции в СВ2 / НСВ2 (0,71), непосредственно за ней следует СВ2 / НСВ1 (0,69) и значительно отстает НСВ1 / НСВ2 (0,42); ту же тенденцию выявляет и среднее значение этого коэффициента, хотя оно менее устойчиво к выбросам: 0,68 – 0,50 и 0,40;
2) высокий (с коэффициентом Спирмена выше 0,7) уровень корреляции на большем количестве случаев (семь из 11) отмечен для СВ2 / НСВ2, для СВ2 / НСВ1 он отмечается в пяти случаях из 11; для НСВ1 / НСВ2 -- в трех случаях, еще в двух -- обратная корреляция;
3) обобщая предшествующие наблюдения, можно сделать вывод о том, что по итогам анализа диахронической частотности на 11 видовых тройках наибольшей согласованностью изменения таковой по обоим параметрам (медиана и среднее по выборке, а также количество случаев с высоким уровнем коэффициента Спирмена) характеризуется пара СВ2 / НСВ2 (две префиксальные формы), сразу за ней следует СВ2 / НСВ1, иную картину мы видим в случае НСВ1 / НСВ2.

5 Выводы и перспективы

Имея в виду риски экстраполяции выводов, сделанных на основе небольшого по охвату материала исследования, на всю языковую систему и не претендуя на окончательность формулировок, приведем несколько более общих соображений, вытекающих из проведенного исследования.

Главный вывод: отказ от широко применяемого в современной лингвистике подхода к решению проблемы словоизменение vs. словообразование через критерии обязательности и регулярности в пользу обращения к количественному подходу посредством учета диахронического изменения частотности словоформ как критерия единства лексического значения (в частности, с использованием GBN) не представляется оправданным в силу целого ряда причин.

I. Прежде всего, это сомнения в обоснованности предложенного подхода дистрибутивной гипотезой [13] и в его предпочтительности по сравнению с использованием критериев обязательности и регулярности, являющихся, в конечном итоге, производной от анализа типов дистрибуции рассматриваемой единицы.
II. Далее, это так называемые ложноположительные и ложноотрицательные результаты применения предложенного в [15, 16] подхода, т. е. наличие случаев, когда явно разные лексемы показывают высокий уровень коэффициента корреляции диахронической частотности (*кот* и *кошка*, ингестивные *съесть, выпить, поесть, попить*), и случаев низкого уровня корреляции словоформ одной лексемы (*был, буду, будет*; *нам* и *нами*; *am going to* и *was going to*).

III. Наконец, наше исследование в рамках предложенного в [15, 16] подхода на материале русских аспектуальных троек показало, что уровень корреляции частотности членов видовой тройки, самим фактом своего вхождения в нее характеризуемых высокой степенью близости лексической семантики, существенно различается при попарном разбиении тройки, см. наблюдения 1-2 выше. Полученные результаты подтверждают трактовку НСВ1 в видовой тройке как «джокера», «выполняющего чужие функции» [1: 106]. С учетом результатов исследования, позволим себе расширить это понятие и утверждать, что «джокером» симплекс НСВ1 выступает не только в случае замещения приставочного НСВ2 (без различий по виду), но и выступая аспектуальным партнером приставочного СВ2. Как показывают результаты исследования изменения частотности, во втором случае (НСВ1 как видовой партнер для СВ2) роль «джокера» симплекс выполняет значительно лучше, чем в первом (см. наблюдения в разделе 4 относительно двоек СВ2 / НСВ1 и НСВ1 / НСВ2), на чем и базируется понятие приставочной видовой пары.

Завершая подведение итогов применения анализа частотности членов видовой тройки в диахронии на материале GBN, отметим, что существенно различные результаты по тройкам (Приложение А) позволяют предположить, что симплексы НСВ1 с неодинаковой легкостью выступают в роли «джокера» как для СВ2, так и для НСВ2, что может быть обусловлено различными факторами, в том числе связанными с лексической семантикой симплекса, степенью его полисемичности, востребованностью в качестве синонима для приставочных СВ2 и НСВ2 в профессиональных языках [18: 248-257], явлением депрефиксации [Ibid: 274] и др. Все это подлежит дальнейшему изучению[12].

Условные обозначения

НСВ – несовершенный вид; СВ – совершенный вид; НСВ1 – симплекс НСВ (типа *лить*), СВ1 – симплекс СВ (типа *дать*); СВ2 – приставочный глагол СВ, дериват НСВ1 (типа *про-лить*); НСВ2 – т. наз. вторичный имперфектив, дериват СВ2, обладающий префиксом и суффиксом имперфективации (типа *про-ли-ва-ть*).

Благодарности

Авторы выражают благодарность Е. В. Еникеевой за помощь в работе с данными НКРЯ.

References

[1] Apresjan Yu. D. Interpretation of redundant aspectual paradigms in the defining dictionary // Apresjan Yu. D. Selected works. Vol. II: Integrated description of language and systematic lexicography [Izbrannye trudy. Vol. II: Integral'noe opisanie yazyka i sistemnaya leksikografiya]. — Moscow: Yazyki Russkoi Kul'tury, 1995. — P. 103–114.

[2] Belikov V. I. (2016), What and how can a linguist get from digitized texts? [Chto i kak mozhet poluchit' lingvist iz ocifrovannyh tekstov?], Siberian Journal of Philology [Sibirskij filologicheskij zhurnal], 3, pp. 17-34.

[3] Bybee J. L., Perkins R., Pagliuca W. (1994), The evolution of grammar: tense, aspect and modality in the languages of the world. Chicago: University of Chicago Press.

[4] Chuikova O. Iu. (2018), The Semantics of Future in Russian, English and Spanish: The Interaction of Temporality, Aspectuality and Modality. Candidate of Philology Thesis [Semantika budushchego vremeni v russkom, anglijskom i ispanskom yazykah (vzaimodejstvie temporal'nosti, aspektual'nosti i modal'nosti): diss. na soiskanie stepeni kand. filol. nauk]. St. Petersburg.

[5] Dahl Ö. (1985), Tense and Aspect Systems. — Oxford: Blackwell.

[6] Dahl Ö. (2000), The grammar of future time reference in European languages // Ö. Dahl (ed.). Tense and Aspect in the Languages of Europe. — Berlin, New York: Mouton de Gruyter. — P. 309–328.

[7] Gorbova E.V. (2017), Aspectual formation of Russian verbs: Inflection, derivation, or a set of quasigrammemes? ("Sore points" of Russian aspectology revisited) [Russkoe vidoobrazovanie: slovoizmenenie, slovoklassifikaciya ili nabor kvazigrammem? (eshche raz o bolevyh tochkah russkoj aspektologii)], Topics in the study of language [Voprosy yazykoznaniya], 1, pp. 24–52.

[12] Отметим также тот факт, что проведение аналогичного исследования тех же 11 аспектуальных троек на таком источнике языкового материала, как НКРЯ (см. Приложение В), дает при общем взгляде на средние величины сходные, однако разнящиеся относительно конкретных троек, результаты.

[8] Gorbova E.V. (2020), Aspectual triplets of the Russian verb in diachrony: Evidence from the Russian National Corpus [Vidovye trojki russkogo glagola v diahronii (na materiale NKRYa)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2020” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2020”], pp. 321-347. DOI: 10.28995/2075-7182-2020-19-321-347
[9] Kasevich V.B. (2019), Problems of Semantics [Problemy semantiki]. — St. Petersburg: St. Petersburg University Press.
[10] Maslov Yu.S. (2004), Selected Works: Aspectology. General Linguistics [Izbrannye trudy: Aspektologiya. Obshchee yazykoznanie]. — Moscow, Yazyki slavyanskoi kul'tury Publ.
[11] Mel'chuk I.A. (1997), Course in General morphology [Kurs obshchej morfologii]. Vol. 1. — Moscow, Vienna: Yazyki russkoj kul'tury, Venskij slavisticheskij al'manah, Izdatel'skaya gruppa «Progress».
[12] Plungian V. A. (2011), Introduction to grammatical semantics: Grammatical meanings and grammatical systems of languages of the world [Vvedenie v grammaticheskuyu semantiku. Grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira]. — Moscow: RGGU, 2011.
[13] Sahlgren, M. (2008), The Distributional Hypothesis. From context to meaning, Distributional models of the lexicon in linguistics and cognitive science [Special issue], Rivista di Linguistica, Vol 20(1), pp. 33–53.
[14] Solovyev V.D., Bochkarev V.V. (2015), Evolution of the aspectual triplets in Russian through the prism of Google Ngram [Evolucija aspectual’nyh tripletov v russkom jazyke cherez prizmu Google Ngram], Proceedings of the international conference «Corpus linguistics – 2015» [Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika – 2015»]. St. Petersburg, SPbSU [SPbGU], pp. 425-434.
[15] Solovyev V.D., Bochkarev V.V. (2022), The case for aspectual pairs reopened [Delo ob aspektual'nyh parah otkryvaetsya vnov'], Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya], Vol. 78, pp. 67–98. DOI: 10.17223/19986645/78/4
[16] Solovyev V., Bochkarev V., Bayrasheva V. (2022), Aspectual pairs: Prefix vs. sufix way o formation, Russian Journal of Linguistics, Vol. 26(4), pp. 1114–1135. DOI: 10.22363/2687-0088-27394.
[17] Zaliznyak A.A. (2002), «Russian nominal inflection» with selected works on Modern Russian and general linguistics [«Russkoe imennoe slovoizmenenie» s prilozheniem izbrannykh rabot po sovremennomu russkomu yazyku i obshchemu yazykoznaniyu]. Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul’tury], 2002.
[18] Zalizniak Anna A., Mikaelyan I.L., Shmelev A.D. (2015), Russian aspectology: In defense of the aspectual pair [Russkaya aspektologiya: v zashchitu vidovoi pary] — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul’tury].

Приложение А. Графики и попарные коэффициенты корреляции в аспектуальных тройках

1) *свариться — вариться / свариваться*

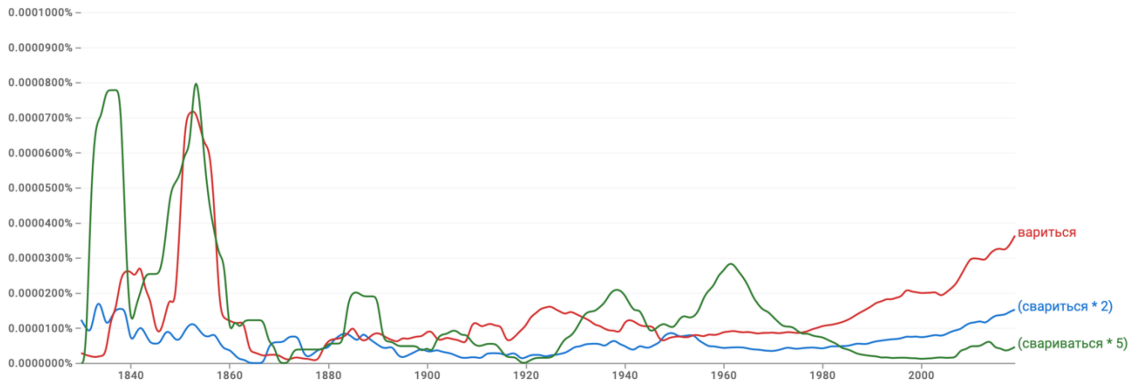

Рисунок А1: Графики частот *вариться – свариться – свариваться*

Пара	коэффициент Спирмена	коэффициент Пирсона
свариться / вариться	0.38	0.45
свариться / свариваться	0.16	0.43
вариться / свариваться	-0.03	0.4

Таблица А1: Попарные коэффициенты корреляции в тройке *свариться — вариться / свариваться*

2) *сварить — варить / сваривать*

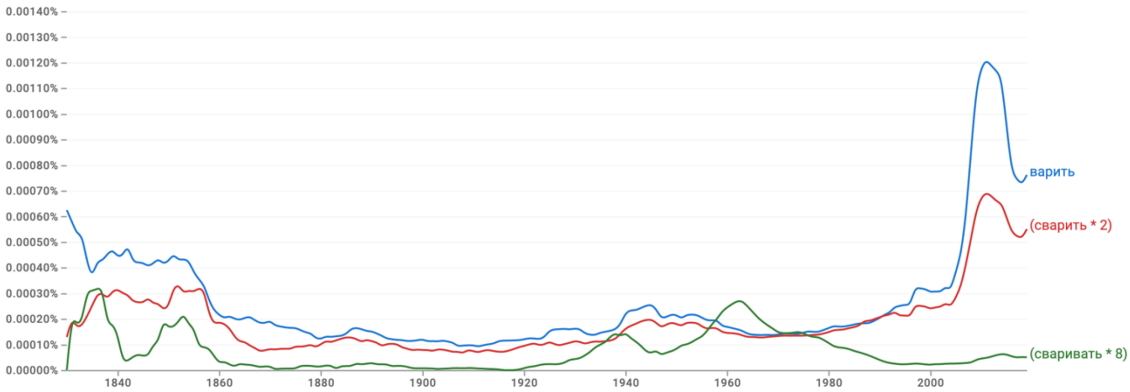

Рисунок А2: Графики частот *варить – сварить – сваривать*

Пара	коэффициент Спирмена	коэффициент Пирсона
сварить / варить	0.89	0.95
сварить / сваривать	0.58	0.14
варить / сваривать	0.44	0.11

Таблица А2: Попарные коэффициенты корреляции в тройке *сварить — варить / сваривать*

3) *съесть — есть / съедать*[13]

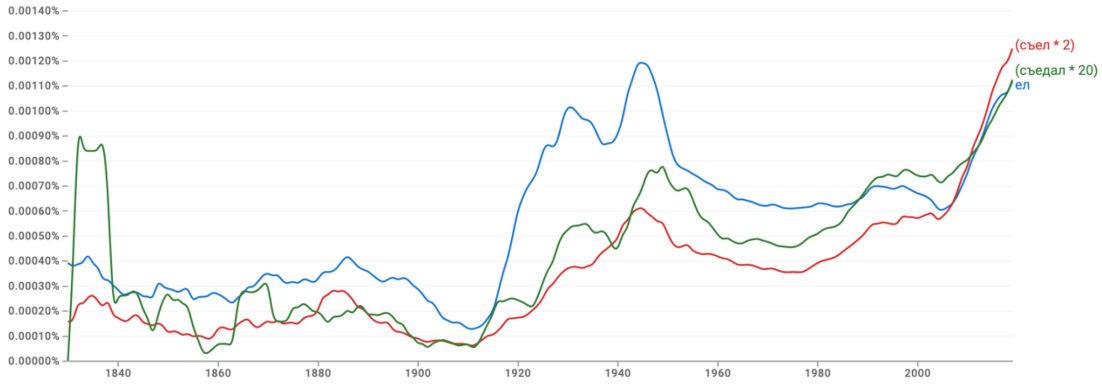

Рисунок А3: Графики частот *ел – съел – съедал*

Пара	коэффициент Спирмена	коэффициент Пирсона
съел / ел	0.87	0.8
съел / съедал	0.88	0.86
ел / съедал	0.77	0.74

Таблица А3: Попарные коэффициенты корреляции в тройке *съесть — есть / съедать*

4) *оторвать — рвать / отрывать*

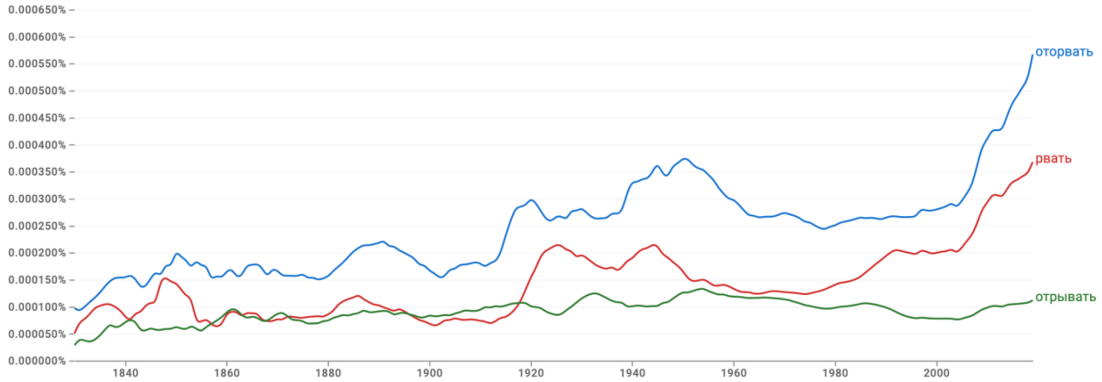

Рисунок А4: Графики частот *рвать – оторвать – отрывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
оторвать / рвать	0.82	0.88
оторвать / отрывать	0.71	0.67
рвать / отрывать	0.41	0.36

Таблица А4: Попарные коэффициенты корреляции в тройке *оторвать — рвать / отрывать*

[13] Для этой тройки рассмотрение инфинитивных форм приводит к некорректным результатам в связи с омонимией инфинитива *есть* и презентной формы глагола *быть*, поэтому исследование проведено на формах претерита.

5) *пробить — бить / пробивать*

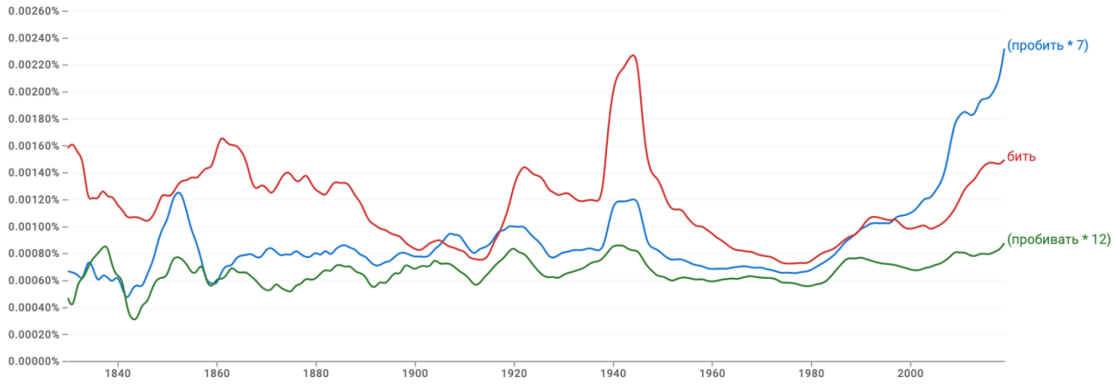

Рисунок А5: Графики частот *бить – пробить – пробивать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
пробить / бить	0.26	0.28
пробить / пробивать	0.71	0.62
бить / пробивать	0.27	0.36

Таблица А5: Попарные коэффициенты корреляции в тройке *пробить — бить / пробивать*

6) *сгореть — гореть / сгорать*

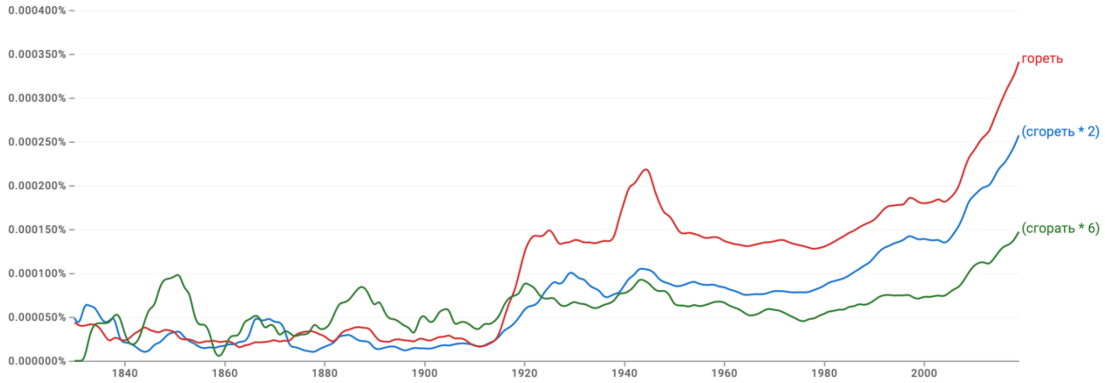

Рисунок А6: Графики частот *гореть – сгореть – сгорать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
сгореть / гореть	0.89	0.95
сгореть / сгорать	0.67	0.72
гореть / сгорать	0.74	0.74

Таблица А6: Попарные коэффициенты корреляции в тройке *сгореть — гореть / сгорать*

7) *сорвать — рвать / срывать*

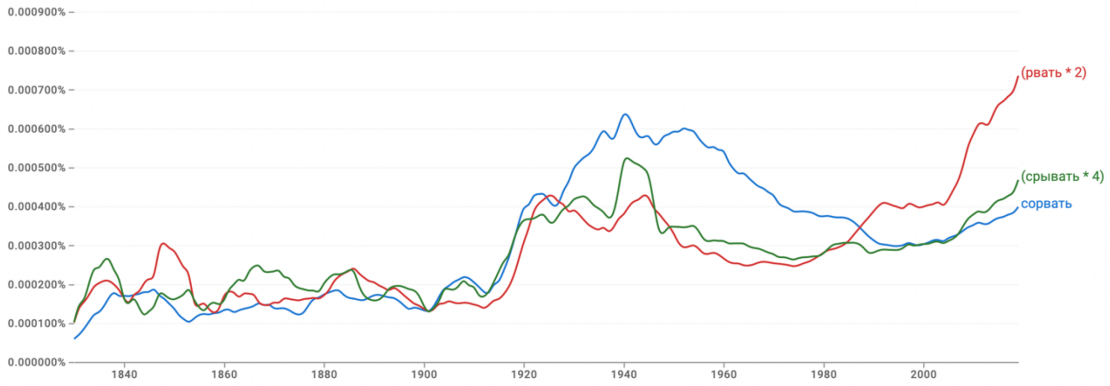

Рисунок А7: Графики частот *рвать – сорвать – срывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
сорвать / рвать	0.69	0.57
сорвать / срывать	0.83	0.85
рвать / срывать	0.84	0.79

Таблица А7: Попарные коэффициенты корреляции в тройке *сорвать — рвать / срывать*

8) *разорвать — рвать / разрывать*

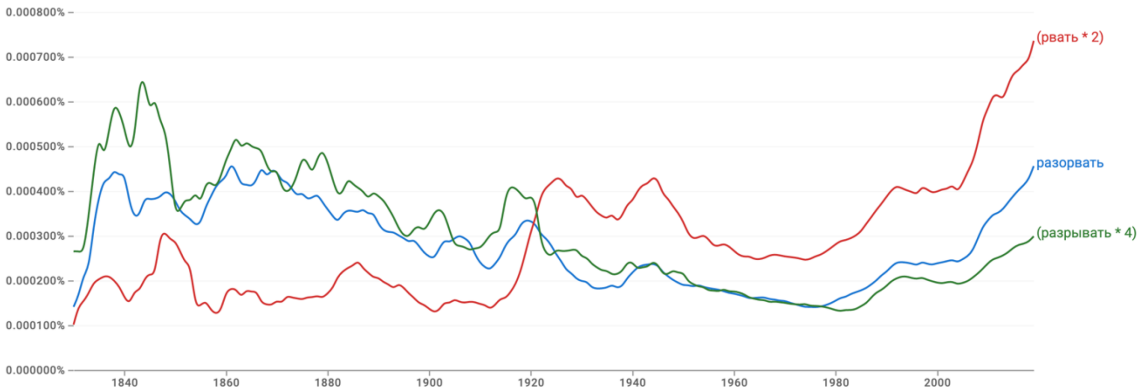

Рисунок А8: Графики частот *рвать – разорвать – разрывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
разорвать / рвать	-0.28	-0.13
разорвать / разрывать	0.88	0.85
рвать / разрывать	-0.53	-0.45

Таблица А8: Попарные коэффициенты корреляции в тройке *разорвать — рвать / разрывать*

9) *разбить — бить / разбивать*

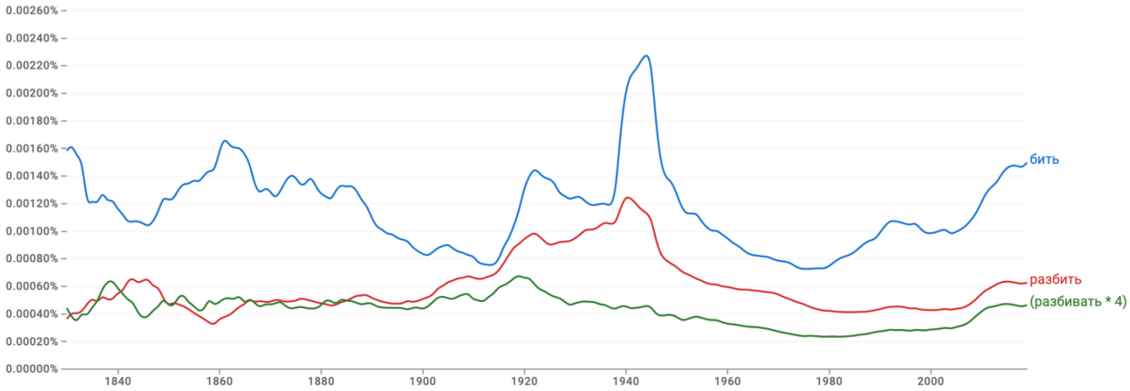

Рисунок А9: Графики частот *бить – разбить – разбивать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
разбить / бить	0.13	0.41
разбить / разбивать	0.42 (0.4199)	0.39 (0.3906)
бить / разбивать	0.42 (0.4181)	0.39 (0.3934)

Таблица А9: Попарные коэффициенты корреляции в тройке *разбить — бить / разбивать*

10) *намазать — мазать / намазывать*

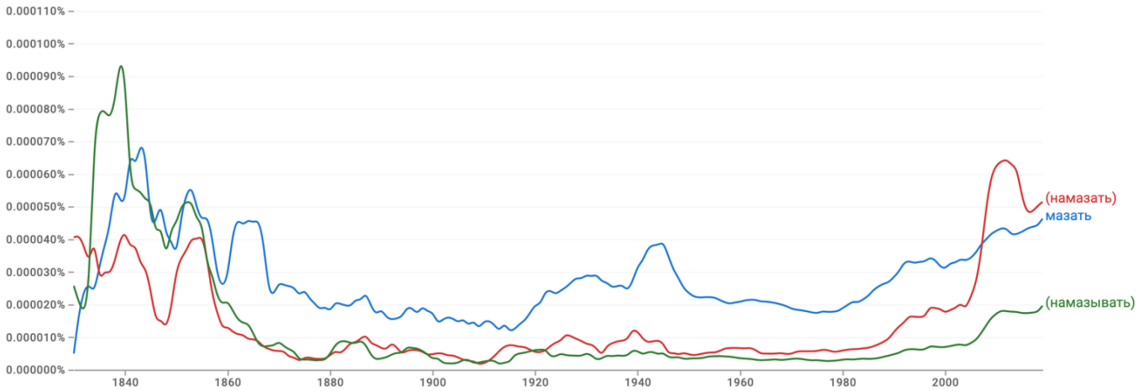

Рисунок А10: Графики частот *мазать – намазать – намазывать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
намазать / мазать	0.73	0.66
намазать / намазывать	0.80	0.60
мазать / намазывать	0.73	0.67

Таблица А10: Попарные коэффициенты корреляции в тройке *намазать — мазать / намазывать*

11) *лить — налить / наливать*

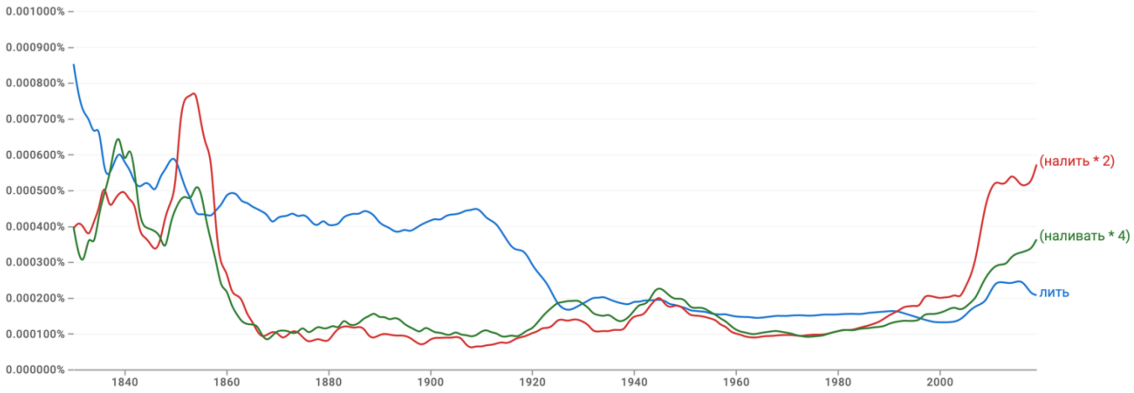

Рисунок А11: Графики частот *лить – налить – наливать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
налить / лить	0.15	0.36
налить / наливать	0.88	0.89
лить / наливать	0.31	0.48

Таблица А11: Попарные коэффициенты корреляции в тройке *налить — лить / наливать*

Приложение Б. Пример кода на Python для расчета коэффициентов корреляции

```
#https://www.geeksforgeeks.org/scrape-google-ngram-viewer-using-python/
import pandas as pd
import scipy.stats as stats
import requests
import urllib

def run_query(query, start_year=1830, end_year=2019,
              corpus='ru-2019', smoothing=3):
    query = urllib.parse.quote(query)
    url = (
        'https://books.google.com/ngrams/json?content=' +
        query + '&year_start=' + str(start_year) + '&year_end=' +
        str(end_year) + '&corpus=' + str(corpus) + '&smoothing=' +
        str(smoothing) + ''
    )

    response = requests.get(url)
    output = response.json()
    return_data = []
    if len(output) == 0:
        return "Нет данных для Ngram."
    else:
        for i in range(len(output)):
            return_data.append(output[i]['timeseries'])
    return return_data[0]

df = pd.DataFrame(
    {
        'год':list(range(1830,2020)),
        'варить':run_query('варить'),
        'сварить':run_query('сварить')
    }
)

df[['варить', 'сварить']] = (
    df[['варить', 'сварить']]
    .apply(lambda x: x*100)
) #переводим значения для всех столбцов, кроме "год", в проценты

# коэффициент корреляции Спирмена
rho, p = stats.spearmanr(df['сварить'], df['варить'])
print(rho, p)

#коэффициент корреляции Пирсона
df['сварить'].corr(df['варить'])
```

Приложение В. Данные дополнительного исследования аспектуальных троек, проведенного по НКРЯ (ruscorpora.ru/chart)

Аспектуальная тройка	СВ2/НСВ1	СВ2/НСВ2	НСВ1/НСВ2
оторвать – рвать/отрывать	0,71	0,59	0,59
съел – ел/съедал	0,71	0,49	0,21
сварить – варить/сваривать	0,70	0,26	0,49
намазать – мазать/намазывать	0,42	0,05	0,34
сгореть – гореть/сгорать	0,41	0,37	0,28
разбить – бить/разбивать	0,29	0,16	0,50
сорвать – рвать/срывать	0,28	0,50	0,51
свариться – вариться/свариваться	0,12	0,09	0,06
налить – лить/наливать	-0,19	0,15	0,08
пробить – бить/пробивать	-0,28	0,75	-0,40
разорвать – рвать/разрывать	-0,28	0,34	-0,23
Среднее	**0,26**	**0,34**	**0,22**
Медиана	**0,29**	**0,34**	**0,28**

Таблица В1: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ1)

Аспектуальная тройка	СВ2/НСВ1	СВ2/НСВ2	НСВ1/НСВ2
пробить – бить/пробивать	-0,28	0,75	-0,40
оторвать – рвать/отрывать	0,71	0,59	0,59
сорвать – рвать/срывать	0,28	0,50	0,51
съел – ел/съедал	0,71	0,49	0,21
сгореть – гореть/сгорать	0,41	0,37	0,28
разорвать – рвать/разрывать	-0,28	0,34	-0,23
сварить – варить/сваривать	0,70	0,26	0,49
разбить – бить/разбивать	0,29	0,16	0,50
налить – лить/наливать	-0,19	0,15	0,08
свариться – вариться/свариваться	0,12	0,09	0,06
намазать – мазать/намазывать	0,42	0,05	0,34
Среднее	**0,26**	**0,34**	**0,22**
Медиана	**0,29**	**0,34**	**0,28**

Таблица В2: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ2

Комментарий к таблицам В1 и В2

Общие средние результаты по Таблицам 2 и 3 основного текста статьи (результаты исследования по GBN) и по Таблицам В1 и В2 (результаты аналогичного исследования по НКРЯ) не совпадают количественно, однако совпадают по парам словоформ: самый высокий коэффициент Спирмена в среднем в парах СВ2 / НСВ2, далее следуют СВ2 / НСВ1, на последнем месте пары НСВ1 / НСВ2. При этом данные по отдельным тройкам, полученные по двум разным источникам материала (GBN и НКРЯ) существенно разнятся (ср. данные из Таблиц 2 и 3 и Таблиц В1 и В2).

		СВ2 / НСВ1	СВ2 / НСВ2	НСВ1 / НСВ2
GBN	среднее	0,50	0,68	0,40
	медиана	0,69	0,71	0,42
НКРЯ	среднее	0,26	0,34	0,22
	медиана	0,29	0,34	0,28

Таблица В3: Сравнение общих средних результатов коэффициентов корреляции Спирмена в тройках (по GBN и по НКРЯ)

Приложение Г. Пример кода на Python для расчета коэффициентов корреляции (НКРЯ)

```
import pandas as pd
import scipy.stats as stats
import requests
import urllib

def run_query_w_smoothing(query, start_year=1830, end_year=2020,
                          smoothing=3):
    query = urllib.parse.quote(query)
    url = (
        'https://processing.ruscorpora.ru/graphic.xml?env=alpha' +
        '&mode=graphic_main&mycorp=&mysent=&mysize=&mysentsize=' +
        '&mydocsize=&dpp=100&spp=&spd=1&text=lexform' +
        '&sort=i_year_created&g=i_year_created&lang=ru&nodia=1&req=' +
        query + ',&startyear=' + str(start_year) + '&endyear=' +
        str(end_year) + '&smoothing=' + str(smoothing) + '&format=json' +
        '&total=2&showChart=false&tableIsRender=false'
    )

    response = requests.get(url)
    output = response.json()
    year_freq_dict = {}
    for i in range(len(output['values'][0]['data'])):
        year_freq_dict[output['values'][0]['data'][i][0]]output[
            'values'][0]['data'][i][1]
    smoothed_data = []
    for year in range(start_year, end_year+1):
        start_smoothing = max(0, year - smoothing)
        end_smoothing = min(end_year, year + smoothing)
        freq_sum = 0
        for y in range(start_smoothing, end_smoothing+1):
            freq_sum += year_freq_dict.get(str(y), 0)
        smoothed_data.append(freq_sum / (2 * smoothing + 1))
    s = pd.Series(smoothed_data, index=range(start_year, end_year+1))
    return s

df = pd.DataFrame(
    {
        'варить': run_query_w_smoothing('варить'),
        'сварить': run_query_w_smoothing('сварить')
    }
)

# коэффициент корреляции Спирмена
rho, p = stats.spearmanr(df['сварить'], df['варить'])
print(rho, p)
#коэффициент корреляции Пирсона
df['сварить'].corr(df['варить'])
```