АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 811.161.1'322.2

М.Ю. Михеев, Л.И. Эрлих

Идиостилевой профиль и определение авторства текста по частотам служебных слов*

Обсуждается возможность составления 100-словного списка русских служебных слов и выражений, т.е. наиболее частых в языке союзов, предлогов, частиц, дискурсивных слов, вводных оборотов, устойчивых наречных конструкций, фразеологизмов, стандартных средств пара- и гипотаксиса, а также наиболее употребительных сочетаний из них. Всякую такую единицу, потенциальный статистический маркер автора, мы условно называем языковой, или текстовой скрепой. Проверяется гипотеза о том, что автора и его текст могут исчерпывающим образом характеризовать как раз не «ключевые» слова, но и слова наименее значимые, наименее заметные, и что при помощи списка из сотни наиболее частотных подобных скреп можно будет находить реального автора текста или, по крайней мере, определять близость исследуемого текста к стилистике того или иного автора, чьи тексты уже имеются в исходной базе.

В идиостилевой профиль должно попасть всё, что есть у писателя наиболее частотного или же, наоборот, редкостного, что отличает одного автора от другого, независимо от того, на какую тему он пишет.

Ключевые слова: *идиостиль, служебные слова, частота слова, национальный корпус русского языка, русская литература*

В информационно-поисковых системах давно известна задача индексирования текстов, с разметкой их – теми словами, которые способны выступать в качестве *ключевых*, описывая сам предмет и тему исследования. Помимо темы, лингвисты умеют выделять еще и *рему* высказывания, а также разнообразные его *презумпции* и *импликации*. Но есть другая задача – распознавание авторства: для нее более важным оказывается не то, **о чем** говорит автор, а то, **как** он это говорит. Точно так же, как пишущего или произносящего мы узнаем по почерку или по голосу (а машинку, на которой напечатан текст, можем узнать по характерным ее дефектам), оказывается, что можно распознать и самого автора – на основании статистического анализа банальных языковых выражений, которыми он пользуется при формулировании своих мыслей, подчас даже не отдавая себе в этом отчет. При этом учет ключевых, полнозначных слов, столь важных при индексировании, оказывается совсем не так эффективен, как подбирание того ставшего крылатым благодаря стихотворению Ахматовой – «словесного сора», каким человек неизбежно уснащает свое высказывание, почти того не замечая.

В лингвостатистике существует 100-словный список М.Сводеша, задающий, как известно, лексику, наименее подверженную изменениям в языке, по которой можно рассчитать скорость синонимических замен базового лексического фонда. Сам список предполагается примерно одинаковым для любых языков и служит как бы "лингвистическими часами"– по изменениям в нем можно определить время распада мертвого языка. Подобно этому мы предлагаем 100-словный список русских служебных слов и выражений, т.е. наиболее часто употребляемых в языке союзов, предлогов, частиц, дискурсивных слов, вводных оборотов, устойчивых наречных конструкций и фразеологизмов, а также употребительных сочетаний из них, заключающих в себе сразу несколько таких единиц. При помощи этого списка, как нам представляется, можно будет находить реального автора текста или, по крайней мере, определять близость исследуемого текста к стилю того автора, чьи тексты уже имеются в исходной базе. За исходную базу взят Национальный корпус русского языка (далее он будет обозначаться сокращенно – **НК**). В ста-

* Работа финансируется грантами РФФИ № 16-06-00070 «Структура многокомпонентных коннекторов русского языка и принципы ее представления в лингвистических базах данных» и № 18-012-00220 А «Создание алгоритма идентификации авторского идиостиля на основании частотности употребления служебных слов» (ФИЦ ИУ РАН).

тье также описывается методика выявления *идиостилевого* (или же *идиолектного*) *профиля* писателя, под которым мы понимаем набор наиболее частых в его текстах, характерных именно для него служебных единиц языка, элементов этого самого 100-словного списка. Для краткости все элементы такого списка будут именоваться просто текстовыми, или языковыми, *скрепами* [1]. Они же выступают и потенциальными *статистическими маркерами* стиля писателя (сокращенно **СМ**), обособляя один идиостиль от другого. Эти единицы охватывают собой как отдельные слова, так и целые синтаксические конструкции, максимально употребительные, «ходовые» у данного автора (его положительные маркеры), но также и – минимально употребительные, отрицательные маркеры [2].

Здесь мы опираемся на высказанную более ста лет назад гипотезу Н.А. Морозова о том, что наиболее показательным для характеристики авторского стиля являются частоты употребления служебных слов [3]. Упомянем еще одну, но уже более проблематичную гипотезу – супругов Т.Г. и В.П. Фоменко, которые в результате исследования текста романа «Тихий Дон» на рубеже 70-х и 80-х гг. прошлого века, поставили под сомнение авторство М.А. Шолохова, придя к выводу, что по крайней мере первая половина этого произведения (до середины 6-й части) по стилистике ближе к текстам Федора Крюкова [4]. Впрочем, надежно подтвердить эту гипотезу статистическими данными, вопреки уверенности авторов, или же ее опровергнуть – до сих пор никому так и не удавалось. Даже подсчеты Гейра Хетсо и его сотрудников, пришедших к совершенно, казалось бы, противоположным выводам [5], окончательного ответа на это не дали. В отличие от обоих подходов мы предлагаем сосредоточиться как на СМ именно на статистике служебных слов, но рассматриваем служебные слова расширительно, т.е. включаем в их число множество дискурсивных слов и вводных конструкций, обстоятельственные и наречные обороты, а также формальные средства стандартных синтаксических средств пара- и гипотаксиса – такие, к примеру, как определительные придаточные с указательным и союзным словом: *тот…, который*, причинные: *потому…, что*, следственные, условные и т.д. и т.п. Предельно расширяя понятие, все скрепы можно было бы назвать «универсальным грамматическим набором» [6].

Интуитивно понятно и легко подтверждаемо на практике, что всякий автор отличается преимущественным употреблением некоторого числа одних слов и их сочетаний и неупотреблением какого-то числа других, чем и выделяется среди большинства современников, предшественников, а также последователей и потомков. Иначе говоря, можно фиксировать набор его ключевых, излюбленных – порой удивляющих читателя, а иногда уже успевших набить оскомину выражений, авторских словечек, языковых клише, его «коньков», сочетаний-стереотипов (даже попросту слов-паразитов в лексиконе). Мы называем подобные выражения для краткости – «скрепами», объединяя вместе под этим понятием союзы, частицы, дискурсивные слова (*ведь, вообще, в целом* и др.), междометия, наречные и предложные группы (*вдруг, ни с того ни с сего, как бы там ни было…*), их устойчивые сочетания и всякую иную неконцептуальную "смазку", скрепляющую смысловые, собственно содержательные куски текста. Их все следует понимать как прямые антиподы *ключевых* слов. Это как бы незаметные, невидимые единицы, неощутимые для носителя языка, или же ощутимые в самой незначительной степени – в сравнении с единицами концептуальными, употребление которых происходит в гораздо большей степени осознанно. Тем самым мы предпринимаем новую попытку определения авторства – по сравнению с альтернативным подходом одного из нас в 2010 г. [7].

Для составления 100-словного списка скреп нами взяты тексты семи следующих русских писателей-прозаиков XIX-XX века, содержащиеся в Национальном корпусе русского языка: Гоголя, Тургенева, Достоевского, Толстого, Бунина, Горького и Набокова; причем их тексты берутся как суммарно (скажем, «весь Гоголь», «весь Тургенев», «весь Достоевский» итд.), так и по отдельности, т.е. по одному, по два наиболее значительных произведения автора, как то: «Мертвые души» у первого, «Дворянское гнездо» и «Отцы и дети» у второго, «Преступление и наказание» и «Братья Карамазовы» у третьего и т.п. В этих подкорпусах (к примеру, во «всем Гоголе» и в «Мертвых душах») в их электронном виде по НК подсчитывается число употреблений всех скреп по 100-словному списку СМ, вычисляются их частоты (в миллионных долях – ipm., или миллипромилле, т. е. 1/10000-й части процента); а для возможности сравнивать между собой, соотнося с неким средним уровнем, эти частоты переведены в проценты. Иным словами, фактически мы сравниваем их доли относительно средней частоты данной скрепы во всем объеме НК. Далее в наборах частот семи избранных авторов определяются экстремумы – максимумы и минимумы для каждой скрепы, т.е. наибольшая частота СМ у кого-то из обсчитанных авторов и частота среди всех прочих наименьшая. Отметим, что таких максимальных и минимальных значений может быть несколько – если сразу у нескольких авторов частоты экстремумов приблизительно совпадают. Остальные частоты, составляющие «средний уровень», могут вообще не учитываться, поскольку только экстремумы, по нашему мнению, составляют отличительные черты авторского стиля. (Впрочем, при расширении нашей задачи оказываются важны не только экстремумы, а и сами значения интервалов частот, однако к этому мы вернемся в конце статьи.)

Итак, в *идиолектный* (или *идиостилевой*) *профиль* должно попасть всё, что есть у писателя наиболее частотного или же, наоборот, редкостного, что и отличает одного пишущего от другого, независимо от того, что, собственно, он пишет... Скажем, слово *Конечно* у Достоевского в романе «Преступление и наказание» встретилось 139 раз (811,87 ipm.), а в «Братьях Карамазовых» 222 раза (754,50 ipm. – при этом последний текст существенно больше по объему): здесь числа в скобках – это частоты, выраженные в миллионных долях. А вот у Льва Толстого во всем романе «Война и мир» то же слово встречается всего лишь 6 раз (соответственно, его частота – 13,30 ipm.) и в «Анне Карениной» – вообще только дважды (т.е. – 7,40 ipm.). Если выразить эти цифры в

относительных частотах – по сравнению со всем корпусом текстов НК (где сейчас около 300 млн. слов), приняв частоту любого слова или сочетания слов в НК за исходный уровень, т.е. 100%, то у Достоевского слово *Конечно* представлено с превышением этой величины (в НК 578,72 ipm.) почти в полтора раза, а у Толстого – с явным дефицитом, измеряемым лишь какими-то единицами процентов. Таким образом мы получаем два важных экстремума употребления данной текстовой единицы – употребление ее с избытком у Достоевского и фактически избегание ее Толстым. Точно так же и по всем остальным потенциальным маркерам, помимо *Конечно*. Если замерить их употребление у наших авторов (для этого у нас есть 100-словный список соответствующих скреп с частотами), у каждого из авторов неизбежно проявятся, станут видны его излюбленные слова-«скрепы», словосочетания, формы преимущественного, или предпочтительного выражения, а также – преимущественного избегания...

В первом приближении мы можем заключить, что *идиостилевой профиль* автора состоит из набора одних только значений частот-максимумов и частот-минимумов на общей 100-словной линейке его скреп. Естественно, что при расширении числа обсчитываемых текстов – и не только самих текстов, но и с увеличением числа писателей, которых мы захотим включить в дальнейшие обсчеты, а также числа СМ, единиц, по которым мы их различаем, сами экстремумы будут раздвигаться, причем в обе стороны: максимумы будут расти, а минимумы стремиться к нулю (у N+1 автора максимальная частота может быть еще выше, чем у N-го, а минимальная – еще меньше). Если же по какой-то из скреп данный писатель пока не попадает ни в лидеры, ни в аутсайдеры, то по какой-то другой (или даже по нескольким другим – пусть не из 100-словного, а из 1000-словного списка, при расширении первого) он имеет возможность туда попасть.

Возражение. И тем не менее, a priori не исключена возможность, что у какого-то автора так и не найдется ни одного потенциального маркера: это можно будет узнать только на большом массиве данных. Действительно, у нового автора должен быть известен, во-первых, достаточно длинный текст X (чтобы мы могли найти в нем хотя бы какое-то число скреп с «наполненными» частотами), во-вторых, он хоть чем-то должен выделяться среди прочих в имеющемся у нас исходном массиве текстов – употреблением либо положительных, либо отрицательных СМ, а главное, не должно оказаться так, чтобы в массиве текстов сравнения на каждый достигнутый максимум или минимум X нашелся бы какой-то автор (и его текст Y), экстремумы которого уже «перекрывали» бы собой экстремумы X-а. Если же последнее условие не выполняется, мы получим попросту нового «бесцветного» автора, чей идиостиль никак не заметен на общем фоне. Но в таком случае мы должны научиться фиксировать, на какой из сравнительного массива текстов его текст более других походит, т.е. устанавливать меру «плагиатичности» X, а уже не меру «языкотворения» его написавшего. (Вот для этой-то, уже следующей задачи и понадобятся интервалы частот, помимо одних экстремумов.)

Серьезная трудность состоит, конечно, и в том, что частота многих скреп не представляет собой постоянной величины, а колеблется от произведения к произведению, причем иногда разброс показателей составляет разы, иногда десятки, иногда – даже сотни раз… Ну, а поскольку мы вынуждены сравнивать между собой интервалы частот, то когда сам интервал достаточно компактен, возможно свести его просто к константе, как, скажем, частоту употребления *Конечно* в двух романах Тургенева, «Дворянском гнезде» и «Отцах и детях»: и здесь и там она попросту одна и та же, 48%, хотя при этом частота в общем, по всему подкорпусу Тургенева значительно выше, 84% (очевидно, за счет малых произведений и «Записок охотника», не включенных в наши подсчеты). Или же, например, частота наречия/вводного слова *Кстати* в целом у Гоголя – 37%, что практически совпадает с частотой и в «Мертвых душах» – 36%.

Порой же – в не слишком большом числе случаев – когда, наоборот, разброс значений по интервалу оказывается слишком велик, мы вынуждены не учитывать показателей частоты, отказываясь от них как от полезной нам реперной точки, не принимать ее во внимание, так как слишком большие колебания не фиксируют частоту. Но в целом таких случаев оказалось не слишком много – менее 15%. Критерием же, или порогом, на котором мы остановились, было избрано более чем двукратное (в 2,5 раза) расхождение частот в текстах одного и того же автора. Так, частота вводного слова *Разумеется* в целом по всем текстам Л.Толстого находится на среднем уровне (таком же, как и в среднем по НК, т.е. близком к 100%), но при этом в отдельных текстах автора она значительно колеблется, подскакивая от «Войны и мира» до «Анны Карениной» – в 5 раз (102: 48-245%; здесь и далее как в самой статье, так и в таблицах первое число перед двоеточием показывает частоту данной скрепы в среднем, в процентах по сравнению с НК, а числа после двоеточия фиксируют, также в процентах, границы интервала частот для этого выражения в конкретных текстах). В данном случае нижняя граница интервала, 48%, указывает уровень *Разумеется* в романе «Война и мир», а верхняя, 245%, – в «Анне Карениной» относительно НК. Чем объяснимы в каждом конкретном случае подобные скачки частот, должен разбираться литературоведческий анализ: вероятнее всего, автором все-таки решается при этом какая-то творческая задача, выводящая употребление скрепы за границы бессознательного..., т.е. данная скрепа в его руках превращается в сознательно применяемое изобразительное средство, и вполне оправданно, что она покидает наш список.

Такой же чрезмерный и потому недопустимый разброс в частотах той же самой скрепы обнаруживаем и в текстах Набокова (65: 41-212% – тоже в 5 раз! в «Лолите» чаще, чем в «Даре»), из-за чего и эти экстремумы приходится отвергнуть. Максимум здесь приходится отдать Достоевскому (соответствующие максимумам числа помечены во всех приводимых таблицах – **п/ж**), а минимум бесспорно принадлежит Бунину (минимумы выделены курсивом; выбракованные же, зашкаливающие показатели частот – подчеркиванием: последние мы будем также называть «желтыми карточками», или «неуравновешенными»

скрепами, такими, отклонения частот которых превышают установленный нами барьер) (табл. 1).

Но вот у Достоевского частота другой скрепы, частицы *Все же*, будучи среди других авторов одной из самых высоких, от «Преступления и наказания» до «Братьев Карамазовых» вырастает более, чем в десять раз, с 15% до 222% (111: 15-**222**%). При таких резких флуктуациях и нестабильности частот очевидно не имеет смысла использовать данные СМ, приходится исключить их из списка максимумов в обоих случаях, приняв за образцы более устойчивые величины: в частности, для *Все же* предпочесть в качестве максимума частоты Набокова, хотя численно они (110: 97-102%) вдвое ниже частоты Достоевского (табл. 2).

В целом же, наверное, наиболее типичными скрепами для художественной прозы следует считать сравнительные союзы *Как бы* и *Как будто*: уровень последнего из СМ в среднем по текстам Достоевского более чем в 4 раза превосходит средний уровень НК, а в романе «Преступление и наказание» частота его выше этого уровня даже более чем в пять раз (409: 189-563%), превышая частоту «Братьев Карамазовых» в 3 раза и опять-таки не удовлетворяя нашему критерию однородности, или «монотонности» частот (2,5 раза) – назовем его так. Совсем недалеко от этих значений отстоят показатели той же скрепы и для текстов Горького (378: 677%). Здесь вначале опять-таки средняя частота, которая почти в 4 раза выше среднего уровня НК, а за ней следует уже не интервал, а частота союза всего в одном романе, «Жизнь Клима Самгина»: та превышает средний уровень НК почти семикратно. Самый же низкий показатель по данной скрепе обнаруживаем у Набокова (161: 82-111%). Причем частота 82%, по его роману «Дар», ниже среднего уровня НК, что для данной скрепы, "фирменной" для художественной прозы, как мы понимаем, несовсем типично.

Похожая ситуация и с союзом или частицей *Как бы*: здесь наибольшие показатели демонстрируют тексты опять-таки Достоевского (283: 358-452%), но самые низкие уже не у Набокова, а сразу у двоих – Толстого (118: 157-170%) и Бунина (118: 183%). Впрочем, не нашлось ни одного автора, у кого частота этой скрепы опускалась бы ниже средней отметки, 100%: в результате минимумы и приходится засчитывать авторам, употребляющим данную скрепу с частотами 120-180% (табл. 3).

Достаточно давно замечен такой маркер текстов Достоевского, типично авторское его словечко, как наречие *Вдруг* [8, 9]. Вот и по нашей таблице из семи авторов легко убедиться, что Достоевский среди всех является бесспорным лидером по частоте его употребления (494: 608-721%). Наивысший показатель – в «Братьях Карамазовых.». Ну, а из остальных авторов ближе всех к нему приближается, во-первых, Бунин (212: 288%), хотя и с весьма заметным отставанием, почти в 200%, что составляет «расстояние» от частоты *Вдруг* в «Жизни Арсеньева» – до частоты Достоевского в целом (последняя же конечно ниже, чем частоты по отдельным его романам); а во-вторых, Тургенев (265: 172-211%), у последнего отрыв частоты по всем его текстам до средней частоты всех текстов Достоевского, как видим, еще выше. Наименее же употребительным (или даже как будто неважным?) словечко *Вдруг* оказывается у двоих – Горького (172: 121%) и Толстого (145: 190-211%). Однако в целом оно также безусловно типично и весьма показательно, и также может считаться «фирменным» для художественной литературы.[1] В частности, ни у кого из семи авторов его частота не падает ниже 100%: самый низкий показатель – в «Жизни Клима Самгина» (172: 121%); правда, в Набоковской «Лолите» он все-таки опускается ниже (202: *113*-244%). Сравним тут и обратное соотношение для двух этих же авторов, Набокова с Горьким, складывающееся по другой скрепе, *Как будто*, где у Набокова также обнаружился минимум (161: *82-111*%). В этом, как и в предыдущем, можно видеть некий антиобщественный, или даже «хулиганский» шаг по отношению к нормам литературы; тогда как у Горького частота данной традиционной для литературы скрепы – одна из максимальных (378: 677%), она используется в его текстах вполне всерьез. А вот Набоков и ее, и *Вдруг* будто специально игнорирует. Впрочем по иным синонимам этого семантического гнезда максимумы принадлежат уже другим авторам: по *Неожиданно* – Бунину (176: 325%), а по *Внезапно* кроме того же Бунина (176: 325%), еще и Тургеневу (600: 338-551%) (табл. 4).

Скрепы, маркирующие такие скорее формально-логические отношения, как причина, следствие и цель, более типичны в ином, научном жанре, а не в художественной прозе, так что неудивительно, что по ним средний уровень НК оказывается в текстах русских писателей-классиков просто недостижим. Но вот по частоте предлога *Благодаря* среди семерых авторов явно выходит вперед наибольший «правдоискатель» – Толстой (хотя даже у него набирается всего лишь 50-60% от общего уровня частоты по НК). Разрыв между экстремумами тут невелик: минимум делят между собой Тургенев, Достоевский и Горький. Положение меняется на симметрично противоположное для очень характерного в литературе маркера, союза/частицы/дискурсивного слова *Ведь*, безусловно имеющего причинно-объяснительное значение, но наряду с этим использующегося еще и для множества разнообразных функций – как то для передачи речевых актов удивления, раздражения, настояния при доказательстве, несогласия и других. Эта скрепа как раз очень типична в текстах Достоевского, где и достигает максимума (303: 317-417); минимальна же – у Толстого (88: 58-75%), разве что еще ниже (ниже даже 50%!) ее частота опускается в снова и снова нарушающей нормы «Лолите» (60: *47-68*%). Малую частоту и здесь можно рассматривать как показатель уклонения автора от принятых литературных канонов, хотя, может быть, и как полусознательное следование Толстому? (табл. 5)

[1] Здесь мы приходим в очевидное противоречие с нашим определением всех скреп – как «словесного сора». Но такова противоречивость самого предмета исследования: именно на таких потенциально значимых для конкретного автора фрагментах текста и пролегает граница того, в чем он способен отдавать себе отчет, а что получается как бы само собой, не проходя через сознание.

Таблица 1

Скрепа // Автор

Скрепа // Автор / число вхождений / частота в ipm. / (частота к **НК**, %)	**Нац. Корпус** весь объем берем за 100%	**Гоголь** весь: 476 187 и «Мертвые души»: 115 143;	**Тургенев** 780 426 и «Отцы и дети»: 54 469 и «Дворянское гнездо»: 46 990	**Достоевский** 1965 416, «Братья К.»: 294 234 и «Преступление и наказание»: 171209	**Л.Толстой** 1926 417, «Анна К»: 270 322 и «Война и мир»: 451296	**Бунин** весь: 633 857 и «Жизнь Арсенева»: 84 122;	**М.Горький** 1826 396 и «Жизнь Клима Самгина»: 560 384;	**Набоков** весь: 722 089 «Лолита»: 101 610 и «Дар»: 104 028
Разумеется	32961 116,29	49 – 102,90 27 – 234,49 (88: 202%)	145 – 185,80 11 – 201,95 4 – 85,12 (160: 73-174%)	952 – 484,38 61 – 207,32 58 – 338,77 (**417: 178-291**%)	228–118,35 77 – 284,85 25 – 55,40 (102: 48-**245**%)	29 – 45,75 7 – 83,21 (*39: 72*%)	230 – 125,93 111 – 198,08 (108: 170-45%)	55 – 76,17 25 – 246,04 5 – 48,06 (65: 41-**212**%)

Таблица 2[2]

Скрепа: Все же

Скрепа	**НацКор**	**Гоголь**	**Тургенев**	**Достоевский**	**Л.Толстой**	**Бунин**	**М.Горький**	**Набоков**
Все же	42939 151,50	26 – 54,60 8 – 69,48 (36: 46%)	40 – 52,25 1 – 18,36 1 – 21,28 (34: 12-14%)	330 – 167,90 99 – 336,47 4 – 23,36 (**111: 15-222%**	31 – 16,09 0 2 – 4,43 (*11: 0-3%*)	76 – 119,90 16 – 190,20 (79: 126%)	93 – 50,92 32 – 58,14 (34: 38%)	120 – 166,18 15 – 147,62 16 – 153,80 **(110: 97-102%)**

Таблица 3

«Фирменные» скрепы художественной литературы

Скрепа	**НацКор**	**Гоголь**	**Тургенев**	**Достоевский**	**Л.Толстой**	**Бунин**	**М.Горький**	**Набоков**
Как будто	52890 186,61	323 –678,30 58 – 503,72 363: 270%)	325 – 416,44 31 – 569,13 16 – 340,50 223: 182-305%)	1500 – 769,20 104 – 353,46 180 – 1051,35 **409: 189-563%**	1202-623,96 172– 636,28 429 –950,60 **334: 341-509%**	197 – 310,80 46 – 546,82 (167: 293%)	1287 – 704,67 708 – 1263,41 **(378: 677%)**	217 – 300,52 21 – 206,67 16 – 153,80 (*161: 82-111%*)
Как бы	80574 337,93	326 – 684,60 93- 807,69 (203:239%)	450 – 576,61 29 – 532,41 19 – 404,34 (171:120-158%)	1879 –956,03 459 –1559,98 207 –1209,05 (**283:358-452**%)	771-400,22 155-573,39 239-529,59 (*118:157-170%*)	252-397,57 52 – 618,15 (*118:183%*)	1088-595,71 490-5824,87 (176:172%)	306 – 423,77 65-639,70 55-528,70 125: 156-189%

[2] Первая строка этой и последующих таблиц повторяет «шапку» из табл.1, уже без указания текстов и их объемов.

Таблица 4

Скрепы: Вдруг, Внезапно, Неожиданно

Скрепа	НацКор	Гоголь	Тургенев	Достоевский	Л.Толстой	Бунин	М.Горький	Набоков
Вдруг	148367 523,47	488-1024,81 118-1024,81 (196:196%)	1083-1387,70 49-899,59 52-1106,62 (265:172-211%)	5084-2586,73 1111-3775,91 545-3183,24 **(494:721-608%)**	1459–757,36 269 – 995,11 499–1105,70 (*145:190-211*%)	703-1109,08 127-1509,71 (212:288%)	1647 – 901,78 356 – 635,28 (*172:121*%)	762 – 1055,27 60 – 590,49 133 – 1278,50 (202:113-244%)
Внезапно	16.048 56,62	9-18,90 1-8,68 (*33:15*%)	265-339,56 17-312,10 9-191,53 (**600: 338-551**%)	116-59,02 32-108,76 11-64,25 (104:192-113%)	10-5,19 2-7,40 2-4,43 (*9:13-8*%)	121-190,89 22-261,52 (**337:462**%)	93-50,92 59-105,28 (90:186%)	120-166,18 14-137,78 12-115,35 (294:243-204%)
Неожиданно	26.938 95,04	13-27,30 7-60,79 (*29:64*%)	44-56,38 4-73,44 5-105,41 (59:77-112%)	163-82,93 28-95,16 15-87,61 (87: 92-100%)	126-65,41 18-66,59 69-152,89 (69:70-161%)	106-167,23 26-309,07 (**176:325**%)	220-120,456 78-139,19 (127:146%)	59-81,71 5-49,21 3-28,84 (*86:52-30*%)

Таблица 5

Скрепы: Благодаря, Ведь

Скрепа	НацКор	Гоголь	Тургенев	Достоевский	Л.Толстой	Бунин	М.Горький	Набоков
Благодаря	29.211- 103,06	16 – 33,60 7 – 60,79 (33: 59%)	14 – 17,94 1 – 18,36 0 (*17: 0-18*%)	32 – 16,28 3 – 10,20 2 – 11,68 (*16: 10-11*%)	114 – 59,18 17- 62,89 23- 50,96 **(57: 49-61%)**	29 – 45,75 5 – 59,43 (44: 58%)	30 – 16,42 7- 12,49 (*16: 12*%)	33 – 45,70 5 – 49,21 6 – 57,67 (44: 48-56%)
Ведь	189189 667,49	499 –1047,91 238 –2066,99 (157: 310%)	1074-1376,17 77-1413,65 54-1149,19 206: 172-212%	3976–2022,98 622 –2113,96 477 –2786,07 **(303: 317-417)**	1135-589,18 136–503,10 176–389,99 (*88: 58-75%*	460 – 727,72 43 – 511,16 (109: 77%)	1661 – 909,44 404 – 720,93 (136: 108%)	291 – 403,00 32 – 314,93 47 – 451,80 (*60: 47-68*%)

Таблица 6

Гнездо скреп возможного, вероятного, предполагаемого

Скрепа	НацКор	Гоголь	Тургенев	Достоевский	Л.Толстой	Бунин	М.Горький	Набоков
Быть может	15471 54,58	2 – 4,20 0- (8: 0%)	103 – 131,98 5 – 91,80 1 – 21,28 242: 39-168%	16 – 8,14 0- 2 – 11,68 (15:0-21%)	6 – 3,11 1 – 3,70 0- *(6: 0-7%)*	6 – 9,47 1 -11,88 (17: 22%)	73 – 39,96 13 – 23,20 (73: 43%)	109 –150,95 7– 68,89 **(277: 126-210%**
Может быть	137355 484,61	383 – 804,31 83 – 720,84 (166: 153%	461 – 590,70 34 –624,21 17 – 361,78 122: 75-129%	2864 – 1457,20 386 – 1311,88 186 – 1086,39 **(301: 224-271%)**	1373-702,34 200–739,86 227–503,00 (145: 104-153%	210 – 331,31 28–332,85 *(68: 69%)*	774–423,79 325 – 579,96 (87: 120%)	355 – 491,62 87-856,21 55 – 528,70 (101: 109-177%
Вероятно	36.787 129,79	38 – 79,47 17 – 147,64 *(61: 114)*	222 – 284,46 11 – 201,95 8 – 170,25 219:131-156%	344 – 175,03 23 – 78,17 35 – 204,43 (135:60-158%	196 – 101,74 33 – 122,08 74 – 163,97 78:94-126%	94 – 148,30 6 – 71,32 *(114: 55%)*	327 – 179,04 212 – 378,31 (138:291%)	212 – 293,59 27–265,72 20 – 192,28 **(226:148-205%**
Вероятнее(й)	1200 4,23	0	3-3,84 1-18,36 0 (91: 434-0%)	51-25,95 12-40,78 3-17,52 (**613:964-414**%)	3-1,56 1-3,70 12,22 (37:87-52%)	0	5-2,74 2-3,57 (65:85%)	0
Возможно	45180 159,40	19 – 39,90 1 – 8,68 (25: 5%)	49 - 62,79 3 – 55,08 2 – 42,56 (39:27-35%)	173 – 88,02 23 – 78,17 23 – 134,34 (55:49-84%)	193–100,19 46 – 170,17 32 – 70,91 **(63: 44-107)**	22 – 34,71 4 – 47,55 *(22: 30%)*	161 – 88,17 91 – 162,39 (**55:102**)	28 – 38,78 10 –98,42 0 24:0-62%
Должно быть	28719 101,33	39 – 81,90 2 – 17,37 (*81:17*%)	157 – 201,17 10 – 183,59 9 – 191,53 (199:181-189%)	293 – 152,10 41 – 139,34 38 – 221,95 (150:138-219%	320 – 166,11 26 – 96,18 44 – 97,50 (164:95-96%)	67 – 105,70 5 – 59,44 (104:59%)	548 – 300,04 241 – 430,06 **(296:424%)**	120 – 166,18 5 – 49,21 17 – 163,42 (164: 49-161%)
Кажется	96661 341,04	252 – 529,20 47 – 408,19 (155: 120%)	513 – 657,33 33 – 605,85 24 – 510,75 (193: 150-178%)	1562 – 794,74 152-516,60 152-887,80 (**233: 151-260**%)	698-362,33 111-410,62 108-239,31 (106:70-120%)	211 – 332,88 21 – 249,64 *(98: 73%)*	941 – 515,22 376 – 670,97 (151:197%)	329 – 455,63 32 – 314,93 64 – 615,22 (134:92-180%)
Очевидно	31.572 111,39	9-18,90 2-17,37 *(17:16%)*	46-58,94 3-55,08 0 (53:49%)	279-141,95 25-84,97 25-146,02 (127:76-131%)	814-422,55 157-580,79 213-471,97 (**379:521-424**%)	51-80,46 10-118,87 (72:107%)	144-78,84 63-112,42 (71:101%)	33-45,70 2-19,68 2-19,23 (41: *18-17*%)
Итого: максимумов минимумов	1386,37	 0 501,91 (*36%*)	 0 0	 829,88 (**60%**) 0	 270,79 (**20%**); 54,58 *(4%)*	 0 1114,84 (**80%**)	 101,33 (**7%**); 0	 184,37 (**13%**); 111,39 (*8%*)

По высказываниям с неуверенностью или с предположением, с вводными словами *Может быть*, лидирует Достоевский (**301: 224-271**%), тогда как более редкую инверсную их форму, *Быть может*, чаще других использует Набоков (**277: 126-210**%); наименее же употребительно *Может быть* у Бунина (*68: 69*%), а *Быть может* – у Толстого (6: *0-7*%). Набоков также лидирует в употреблении *Вероятно* (**226: 148-205**%), аутсайдерами здесь выступают Гоголь и Бунин: частоты обоих – между 55 и 115% (тогда как у Достоевского частота этой скрепы зашкаливает, превышая критерий монотонности). У тех же Гоголя и Бунина оказываются минимумы еще и в употреблении *Возможно*; впереди же по ней сразу двое – Толстой (**63: 44-107**%) и Горький (**55: 102**%). Для Горького весьма характерно предположение с модальностью долженствования, или констатацией искомого, должного положения вещей: ему принадлежит максимум по *Должно быть* (**296: 424**%), тогда как минимум здесь у Гоголя (*81: 17*%). По *Очевидно* впереди всех Толстой (**379:521-424**%), позади же – Гоголь и Набоков, с почти идентичными показателями минимумов (*16-18*%). Вот как выглядит в целом это гнездо скреп (табл. 6).

Итак, в этом гнезде по числу максимумов и суммарному объему частот в них (измеряемому все в тех же ipm. и в % от НК) среди семи авторов лидирует Достоевский, у него 3 максимума, с 60% объема всего гнезда в НК, и ни единого минимума; за ним следует Толстой – с 2 максимумами (и 20% объема) против лишь одного минимума (4%); вслед за ними, на третьем месте, Набоков – с двумя максимумами (это в сумме 13%) и только одним минимумом (8%), далее (или вровень с ним?) Горький – у него только один максимум (7%), но зато вообще нет ни единого минимума; потом Тургенев, у которого нет ни максимумов, ни минимумов; далее Гоголь и Бунин: у каждого из которых только по четыре минимума и ни одного максимума, однако у первого суммарный объем всех минимумов составляет 501,91 ipm. (или 36% гнезда), а у второго значительно больше – 1114,84 ipm. (т.е. 80%)! Как представляется, можно измерять насыщенность текста автора словами, выражающими определенные семантические отношения, определяя таким образом бессознательные авторские предпочтения. На данном примере мы видим, что при выражении возможного и предполагаемого среди семерых безусловно лидирует именно Достоевский, а Бунин, наоборот, к выражению этого типа отношений достаточно равнодушен (или даже враждебен?) – сравним его с Тургеневым, чьи статистические показатели при выражении этих отношений демонстрируют действительно как будто полное «спокойствие», или же гармоничное следование норме.

ВЫВОДЫ

Среди проанализированных авторов можно выделить, с одной стороны, тех, кто владеет более «агрессивным» дискурсом: это писатели, которые активно эксплуатируют собственные авторские предпочтения, внедряя в речь ключевые, или «ударные» словечки и выражения, стремясь как бы изменить, деформировать сам язык (Толстой, Достоевский, Гоголь), а с другой стороны тех, у кого дискурс скорее пассивно-традиционный, выделяющийся, наоборот, более **не**-употребительностью и **не**-поддержкой каких-либо и чем-либо выделяющихся, «ключевых» (для кого-то другого) слов и выражений, т.е. это группа писателей-«охранителей» языка, придерживающихся литературной нормы, не меняющих ее, а ей следующих, утверждающих ее в своих текстах (таковы Бунин, Тургенев, и, как ни странно, Набоков).

Итак, Бунин, Тургенев и Набоков находятся на одном конце шкалы, а Толстой, Достоевский и Гоголь – на противоположном. У первых двоих всего по 4 (или 4+1) максимума, что почти на порядок меньше, чем у последних. У Гоголя – наибольшее число экстремумов (50), рекордсмен же по числу максимумов – Достоевский: у него их 40+9 и при этом всего только 3 минимума – меньше всех! Больше всех минимумов – у Горького (37). Если расположить авторов по числу экстремумов, получим следующую таблицу (знак + означает учет в ней еще и «зашкаливающих» показателей для максимумов и минимумов, например, таких трех максимумов для *Кроме* у Набокова (21+2) как *Меж тем* и *Так_что*) (табл. 7).

Таблица 7

Автор	Кол-во всех экстремумов	Кол-во максимумов	Кол-во минимумов	Кол-во всех «пустых» скреп	Кол-во «желтых карточек»
Гоголь	**50**	29+2	21+1	69	*9*
Горький	48	11+3	**37**	73	*7*
Л.Толстой	47	29+6	18	*57*	**22**
Достоевский	43	**40+9**	*3*	66	**19**
Набоков	33	21+2	12+3	62	**26**
Тургенев	*34*	*4*	30	**77**	**19**
Бунин	34	4+1	30	79	*7*

Таким образом, наиболее «инертными» авторами, не использующими те или иные выражения в качестве СМ, чей язык по этим параметрам почти не подвержен изменениям, оказываются Бунин и Тургенев: у них примерно одинаковое число «пустых» клеток, или таких, где нет ни максимумов, ни минимумов, но у Бунина при этом значительно меньше «желтых карточек» (частот, нарушающих монотонность). Набоков – в своем следовании мейнстриму, или средним частотам употребления выражений-скреп, приближается к тому же «охранительному» полюсу, т.е. отрицательному полюсу языковых инноваций, что и Бунин с Тургеневым, а вовсе не так близок к Толстому, любимому им (о чем он, как известно, громогласно заявлял в своих «Лекциях о русской литературе», читанных в американских университетах), но совсем не далек оказывается и от напоказ поносимого им в этих же лекциях – Достоевского [10]. Но у Набокова еще и рекорд по «желтым карточкам» (26):

их у него даже больше, чем у Толстого (и тем более чем у Достоевского и Тургенева). Бунин и Горький, напротив, рекордсмены как раз по невыходу частот за установленные границы – у каждого из них таких скреп лишь по 7, их частоты среди остальных как раз наиболее однородны. На этом основании Горького также следовало бы причислить к «охранителям» языка, хотя у него же оказывается и наибольшее число минимумов. При этом Гоголь – будучи рекордсменом по числу всех экстремумов, – близок к Бунину и Горькому по малому числу «желтых карточек» (9). У Достоевского же, как было сказано, более всех максимумов, а Толстой выделяется среди остальных, кроме близкого к рекордному количества «выбракованных» скреп (22), еще и тем, что у него наименьшее число «пустых» клеток в таблице (57), таких, которые ничем не выделяли бы его на общем фоне частот НК.

Ну, а теперь, подходя уже к заключению: как можно представить основные расхождения в частотах употребления скреп двух главных «стилесозидающих» или даже, можно сказать: «стилепредержащих» писателей XIX века, Достоевского и Толстого? Их максимумы совпали по 9 скрепам (две из которых «вкладываются» друг в друга). Вот их список по возрастанию значимости, или «веса» по НК, в ipm., округленно: 10 *Для того (...), чтобы*; 12 *Когда (...), тогда*; 82 *Потому (...), что*; 187 *Как будто*; 241 *Так_как*; 560 *То (...), что*; 599 *Потому что*; 809 *Потому*; 2687 *Только*. Т.е. их общий вес (за исключением вкладывающихся) – 4505 ipm. При этом противоположны у них экстремумы (когда у одного – максимум частоты, у другого по ней же – минимум) по 11 скрепам: общий вес таких скреп 3123 ipm. В основном все максимумы тут, конечно, у Достоевского, за исключением только трех, помеченных ниже прямым шрифтом: 49 *Так сказать*; 56 *В целом*; 57 Внезапно; 103 Благодаря; 103 Для того чтобы; 152 *Все же*; 338 *Как бы*; 496 *Лишь*; 523 *Вдруг*; 579 *Конечно*; 667 *Ведь*. Весьма примечательно, что в целом противостояние двух корифеев русской литературы, или число скреп, в которых их расхождения либо максимальны, либо минимальны (но либо тот, либо другой чем-то да выделяется среди собратьев-писателей), занимает собой около 67% всего 100-словного списка, или 86 строк из 127 скреп.

И уже в самом финале: как из этих показателей частот составить идиостилевой профиль и понять, была ли, скажем, в самом деле первая половина «Тихого Дона» написана автором с идиолектом Крюкова, или ли же скорее – Шолохова? Вся методика определения этого, собственно говоря, нами уже изложена: берутся массивы текстов, во-первых, заведомо Крюковский, во-вторых, заведомо Шолоховский – у них через интервалы частот определяются максимумы и минимумы, затем у спорного текста, т.е. первой части «Тихого Дона», в свою очередь, определяются его максимумы, минимумы и интервалы частот у всех скреп, которые можно в нем обнаружить. Наконец, эти три профиля сопоставляются друг с другом, вернее профиль исследуемого текста накладывается на профили первого и второго: в какой из них он легче «входит», за тем и следует признать авторство.

По 100-словному списку (именно, по 127 скрепам), данная процедура была произведена. В результате релевантными, с ненулевыми показателями, оказались 97 скреп, из них по 65 скрепам текст первой половины романа может быть отнесен к перу Михаила Шолохова, а по 32 к перу Федора Крюкова, что составляет, соответственно, 67% и 33%. Или если различать, с одной стороны, точные попадания в интервалы того и другого писателей, а с другой, еще и попадания в результате логического вывода (или «близости» к одному из интервалов), то в пользу авторства Шолохова указывает 77% всех релевантных скреп, а в пользу Крюкова – 23%. Это не отменяет конечно вариант «кражи замысла» или даже кропотливого «переписывания чужого текста», переложения его своим языком, на своем, «вёшенском», а не «глазуновском» диалекте [11], но вариант украденной рукописи с последующим «списыванием из оригинала», причем с ошибками [12], этим отметается, и надеемся, что все-таки достаточно надежно.

* * *

Авторы благодарят за замечания, высказанные в ходе написания этой статьи – С.М. Евграфову, О.Ю. Инькову-Манзотти, Е.Б. Козеренко, М.В. Копотева, Г.Е. Крейдлина, С.Л. Николаева, Н.В. Перцова, Е.В. Рахилину, О.А.Смирницкую, Н.В. Сомина.

СПИСОК ЛИТЕРАТУРЫ

1. Прияткина А.Ф. Текстовые скрепы и «скрепы-фразы» (О расширении категории служебных единиц русского языка) // Русский синтаксис в грамматическом аспекте (синтаксические связи и синтаксические конструкции). – Владивосток: Изд-во Дальневост. ун-та, 2007. – С. 334–344.
2. Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь языка Ф.М. Достоевского. – М.: Языки славянских культур, 2003.
3. Морозов Н.А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд // Известия Отдела русского языка и словесности Императорской Академии наук. – 1915. – Т. 20, кн. 4. – С. 93–127.
4. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Приложение: Кто был автором "Тихого Дона"? // Сб. "Методы количественного анализа текстов нарративных источников". – М.: АН СССР, Ин-т Истории СССР, 1983. – С. 86-109.
5. Хетсо Г., Густавссон С., Бекман Б. Кто написал «Тихий Дон»?: (проблема авторства «Тихого Дона»). – М.: Книга, 1989. – 192 с.
6. Плунгян В.А. Универсальный грамматический набор как инструмент грамматической типологии // Международная конференция, посвященная 50-летию Петербургской типологической школы: Материалы и тезисы докладов. – СПб., 2011. – С. 142-145.

7. Михеев М.Ю. Имплицитное или сокрытое из текста: «Тихий Дон» – М. Шолохова или Ф. Крюкова? // Текст и подтекст. Поэтика эксплицитного и имплицитного. Материалы Международной научной конференции 20-22 мая 2010 г. [в ИРЯ РАН]. – М.: Азбуковник, 2011. – С. 421-429.
8. Гроссман Л. «Вдруг у Достоевского» // Книга и революция, 1921, кн. XX.
9. Белкин А.С. «Вдруг» и «слишком» в художественной системе Достоевского // Читая Достоевского и Чехова. Статьи и разборы. – М., 1973.
10. Набоков В.В. Лекции по русской литературе. – М.: Независимая газета, 1996. – С. 175-218.
11. Михеев М.Ю. О случайных и неслучайных совпадениях – в текстах Ф. Крюкова и М. Шолохова // Логический анализ языка. Адресация дискурса. – М.: Индрик, 2012. – С. 326-42. – URL: http://lit.lib.ru/m/miheew_m_j/text_0010.shtml
12. Чернов А.Ю. Как сперли ворованный воздух. Заметки о «Тихом Доне». – на сайте Несториана. – URL: http://chernov-trezin.narod.ru/ TitulSholohov.htm.

Материал поступил в редакцию 02.11.17.

Сведения об авторах

МИХЕЕВ Михаил Юрьевич – доктор филологических наук, ведущий научный сотрудник НИВЦ МГУ, Институт проблем информатики Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва
e-mail: mihej57@yandex.ru)

ЭРЛИХ Лев Исаакович – ведущий программист НИВЦ МГУ
e-mail: levehr@yandex.ru)