Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2017»

Москва, 31 мая — 3 июня 2017

ЧАСТОТА СЛУЖЕБНЫХ СЛОВ КАК РАЗЛИЧИТЕЛЬНЫЙ ПРИЗНАК *ИДИОСТИЛЯ* (В СВЯЗИ С ГИПОТЕЗОЙ СУПРУГОВ ФОМЕНКО)[1]

Михеев М. (mihej57@yandex.ru)
НИВЦ МГУ; ФИЦ ИУ РАН, Москва, Россия

Эрлих Л. (levehr@yandex.ru)
НИВЦ МГУ, Москва, Россия

Рассматривается гипотеза, высказанная в 80-х годах XX века, о том, что наиболее удобным *стилеметрическим* инструментом при определении авторства текста могут быть частоты служебных слов (многие считали, что использованием этого факта можно опровергнуть статистические выводы известной работы Г. Хьетсо и др. о том, что автором первой половины романа «Тихий Дон» с большей достоверностью является Шолохов, нежели Крюков). В настоящей работе эта начальная гипотеза несколько видоизменена и доработана: сам корпус известных текстов Крюкова за это время серьезно расширен, вместо совокупной частоты всех слов, использовавшейся первоначально, предлагается более детальная методика, описывающая частоты каждого из этих элементов (коннекторов, предлогов, наречий, модальных, дискурсивных и вводных слов) как конкретную составляющую «идиолектного профиля» автора. И хотя на том же списке обсчитанных 54 слов первоначальный вывод авторов XX века вроде бы подтверждается, но перед исследователем XXI-го века встают проблемы при объяснении этого парадоксального результата, рождая сомнения в том, не будет ли финальная картина после расширения списка совсем иной.

Ключевые слова: идиостиль, служебные слова, предлоги, частицы, союзы, модальные выражения, дискурсивные слова, вводные конструкции, разрывные многосоставные коннекторы, статистический анализ, стилеметрия, авторский идиолектный профиль

---

[1] Работа поддержана грантом РФФИ 16-06-00070 «Структура многокомпонентных коннекторов русского языка и принципы ее представления в лингвистических базах данных».

CONNECTORS FREQUENCIES AS A DISTINCTIVE SIGN OF THE INDIVIDUAL STYLE (IN VIEW OF THE COUPLE FOMENKO HYPOTHESIS)

Miheev M. (mihej57@yandex.ru)
Research computing center of Moscow University, Institute of informatics problems; FRC CSC RAS, Moscow, Russia

Ehrlich L. (levehr@yandex.ru)
Research computing center of Moscow University, Moscow, Russia

We study a hypothesis, that had been already expressed in the 1980's of the XX-th century, that the best measuring style tool to determinate and identify the real author of the text can be the connectors frequencies. Many of the linguists believed that the application of this tool would let them to contest stylistic conclusions that G. Kjetsaa and others had presented in his famous woks, stating that it was more Sholokhov than Kryukov who had wrote the novel *And Quiet Flows the Don.* In this article, we modified and developed this initial hypothesis: we enlarged the number of Kryukov texts and instead of the whole frequency of all the words, that were initially used, we offered a more detailed and genuine method that describes frequencies of each of function words (connectors, prepositions, particles, adverbial, modal, parenthetical, discursive words etc.) as a concrete component of the author's individual style. Although the initial conclusion of those authors, based on fifty-four words taken from the same list, seems to be relevant, a researcher of the XXI-st century would probably face the new problems and the new serious doubts, when he will enlarge his connectors list.

Key words: individual style, function words, connectors, prepositions, particles, modal idioms, discursive words, introductive words, complex connectors, statistical analysis, style measuring, author individual profile

«Из мелкой сволочи вербую рать»...
(Пушкин. Домик в Коломне)

§1. Слух

В течение почти 90 лет, с 1928-го года по сей день, имеет хождение следующее устойчивое мнение — что роман «Тихий Дон» написал не Михаил Шолохов, а кто-то другой. Из множества предполагаемых исследователями этого вопроса претендентов на авторство наиболее вероятным считается Федор Дмитриевич Крюков (1870–1920), автор более сотни очерков, рассказов и повестей

(за 30 лет своей творческой деятельности, с 1890 по 1920). В основном они появлялись в газетах и журналах — «Донская речь», столичных «Северном вестнике», «Историческом вестнике», «Сыне Отечества», «Русских Ведомостях», «Речи», а также в журнале «Русское богатство», издававшемся В. Г. Короленко. После того, как Крюков примкнул к белому движению на Дону, в 1918–1919 гг. он печатался в Новочеркасских «Донских Ведомостях» (избранный секретарем Войскового круга, был в эти годы главным редактором этой газеты — официального органа тогдашней Донской республики)[2].

Основные предположения, высказывавшиеся в пользу авторства или, по крайней мере, участия Крюкова (далее сокращенно: К) в написании романа «Тихий Дон» (далее: ТД), состояли в том, что:

- либо он сам написал весь роман, а Шолохов (далее сокращенно Ш), как умел, с ошибками, по утверждению некоторых, его переписал, готовя рукопись для комиссии, которая разбирала дело Ш по обвинению в плагиате: в результате в конце марта 1929 г. комиссия отвергла слухи о плагиате как «злостную клевету, распространяемую врагами диктатуры пролетариата» [2];
- либо что К написал лишь некий протограф романа, а Ш этот текст взял за основу и, существенно переработав и дополнив, так сказать, переписав своим языком, издал под своим именем;
- либо что роман был написан с участием сразу нескольких авторов: текст К послужил лишь начальной точкой для их работы…

Перечисленные выше три версии (на самом деле их больше) подкреплялись различными аргументами, их отстаивали, оспаривали, доказывали разные «неофициальные» исследователи. Так, например, А. Ю. Чернов до сих пор придерживается первой из них [3]; один из авторов данной статьи — сторонник второй; а ростовский историк А. В. Венков высказал третью (что после К текст романа активно редактировался и дорабатывался еще и А. С. Серафимовичем) [4].

Сомнение в авторстве Ш наиболее весомо было заявлено в работе, опубликованной сначала инкогнито «литературоведом Д» (в действительности же И. Н. Медведевой-Томашевской) в книге, изданной в 1974 г. А. И. Солженицыным с предисловием последнего [5]. Гипотезу «Д.» и Солженицына попытались опровергнуть методами компьютерного анализа скандинавские лингвисты под руководством Г. Хьетсо [6], однако статистические показатели, использованные в их работе (сравнение длины предложений, средней длины слов, богатства словарного запаса, долей слов, встречающихся только по одному разу в текстах Ш и К итп.) показались сомнительными и многих не убедили [7].[3]

Появлялись и в последнее время версии о совсем ином составе участников «проекта „Тихий Дон“», уже без К: например, о *Викторе Севском* (настоящее имя: Вениамин Алексеевич Краснушкин) как авторе этой «советской „Войны и мира“» — в книге Зеева Бар-Селлы (Владимира Назарова) [9]. Он считал

---

[2] Статьи этого его последнего периода собраны в книге [1].

[3] В статье [8] отмечено заметное возрастание доли прямой речи в текстах второй половины ТД — как и в целом в текстах Ш, по сравнению с первой половиной ТД.

вероятным привлечение при написании «Они сражались за родину» в качестве литературного негра еще и Андрея Платонова. Допускается также существование некоего безымянного писателя, якобы сидевшего в застенках ОГПУ–НКВД и оттуда поставлявшего своим тюремщикам произведения «красного Льва Толстого»...

Тем не менее, доказать ни одну из «неофициальных» версий с какой-то надежностью никому пока не удавалось. Не удавалось — и опровергнуть...

§2. Гипотеза

Наряду с упомянутой работой Хьетсо статистическим сравнением ТД с текстами Ш и К занимались родители академика-математика А. Т. Фоменко: Тимофей Григорьевич (1910–1992), инженер в углерудодобывающей промышленности, кандидат технических наук, и его жена Валентина Поликарповна (1918–2009), по образованию филолог (работала как учитель русского языка и литературы)[4]. Далее мы будем ссылаться на их работы, используя сокращение ФиФ. В этих работах была выдвинута гипотеза о существовании *авторского инварианта* — в виде процента служебных слов, т. е. процентного отношения трех классов частей речи, *предлогов, союзов* и *частиц* к общему числу слов в тексте.[5] Согласно подсчетам выходило, что в частях 1–5 и на «первых 100 страницах» 6-й части **ТД** (автором которых предположительно мог быть К) этот показатель составляет **19,55 %,** а уже во второй половине романа, ранних «Донских рассказах» (**ДР**), «Поднятой целине» (**ПЦ**), «Они сражались за родину» (**ОсР**) и поздних рассказах, очерках и повестях **Ш**, вместе взятых (т. е. «несомненно Ш»), он на три или даже на пять процентов выше: точнее, процентные показатели лежат в интервале от **22,5** до **24,4**. (Это назовем *выводом*-1 из их гипотезы: он будет ниже нами оспорен.)

Тут следует оговориться, что во время написания работы ФиФ, в 1970–80 гг., тексты К были малодоступны и авторы смогли обсчитать лишь четыре далеко не самых больших, сравнительно ранних его рассказа: «Шаг на месте» (1907), «Жажда» (1908), «Мать» и «Полчаса» (оба последние — 1910). В них около

---

[4] Их фотографию можно видеть на сайте: http://novejshaaj.mybb.ru/viewtopic.php?id=1050 — и там же прочесть воспоминания старшего Фоменко «У подножия». В разделе «Полный список моих научных публикаций» в конце них значатся следующие две работы, одна за 1977 год: В. П. Фоменко, Т. Г. Фоменко Спектральный анализ литературных текстов. (Кто автор Тихого Дона?). 2,5 п. л.; и вторая, за 1990 год: В. П. Фоменко, Т. Г. Фоменко Анализ литературных текстов М. А. Шолохова. 1,5 п. л.

[5] По-видимому, во всех случаях имелся в виду суммарный процент всех этих трех частей речи: по крайней мере, так разумнее всего понимать их работы, хотя это нигде авторами явно не оговаривается. Что такое «инвариант» и каков его смысл, поясняет в комментарии к работе родителей их сын, А. Т. Фоменко: «Главным результатом настоящей работы является обнаружение „авторского инварианта“ для русских литературных текстов. Он позволяет различать некоторых авторов и оказывается полезным при решении проблем, связанных с плагиатом». Здесь и ниже цитирую их работу по электронному тексту из библиотеки Максима Мошкова [10]. Сходный метод установления авторства использует автор системы «Дельта» [15].

42,4 тыс. слов, что составляет менее пятнадцатой части известных и опубликованных на сегодня его текстов. При подсчетах ФиФ процент служебных слов К оказался равен **21** %. Из чего авторы сделали естественный в их логике вывод, что «стилеметрия Крюкова не так уж сильно отличается от „Тихого Дона“», т. е. что стиль **К**, на самом деле, значительно ближе к **ТД**, чем стиль **Ш**. Это можно назвать следствием из вывода-1, однако сам вывод был сделан все-таки на основании неполных данных, что честно признавали сами авторы. Приведем список служебных слов, по которым производились их подсчеты (они перечислены в их работе):

- во-первых, это 24 **предлога**: *в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под*;
- кроме того, 14 **союзов**: *и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто*;
- и наконец — 17 **частиц**: *не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо*. Всего — 55 служебных слов; но впоследствии из числа частиц (!) было исключено дефисное написание *то-есть*, по-видимому, как устаревшее, хотя это специально не оговаривалось.[6]

Несмотря на то что данный список служебных слов неполон, он оказался, как утверждалось авторами, вполне достаточным для различения тех писателей, тексты которых с помощью их методики обсчитывались: в подсчеты было включено около трех десятков писателей — от Фонвизина и Новико́ва до Шолохова и Фадеева. Это можно назвать еще одним «сильным» утверждением, или *выводом*-2 из исходной гипотезы. Не что-то другое, а именно процент служебных слов оказался наиболее различительным параметром из всех остальных показателей, испытанных на ту же роль (как то средней длины слов, средней длины предложений, состава частей речи на первых или на последних местах предложения итд. итп.)[7].

Так, например, согласно данным ФиФ у текстов М. Горького процент служебных слов колеблется в очень узких пределах — всего лишь от 22,0 % до 22,2 %! Однако по нашим обсчетам (мы взяли 6 текстов этого автора, проверив в них частоты служебных слов по их списку), процент получился иной и вообще не пересекающийся с полученным у них, хотя его колебания оказались примерно такими же: от 20,42 до 21,69 %. У других авторов данный параметр, по их данным, также весьма стабилен: так, у Гоголя он составлял 23,5–23,9 % (по нашим данным, на трех текстах, сам процент выше и его колебания более широкие — от 24,01 до 25,37); у Гончарова, согласно ФиФ, 24,9–25,5 % — вот только здесь наши данные, полученные на 4 текстах, наиболее точно совпали с их результатами: от 24,56 до 25,24). Для Достоевского у них — 25,2–25,4 %,

---

[6] По данным НК на сегодня оно встречается, например, в «Идиоте» 160 раз, а «Войне и мире» 1 раз, в целом же абсолютная частота его — 913, или 3,44 в миллионных долях (ipm., instance per million).

[7] По сути дела, это те же статистические параметры, которые рассмотрены и в работе Хьетсо и др.

и наши данные здесь также довольно близки — от 26,42 до 26,57)[8]. Но вот у Лескова, по ним, 25,8–26,2 % — а по нашим обсчетам (в 6 текстах) колебания значительно выше, от 24,16 до 28,45…

Это следовало бы назвать уже, наверное — *выводом*-3 из гипотезы ФиФ, т. е. что «процентная норма» у каждого автора своя собственная, причем колебания ее не выходят за границы 1 %! — но при ближайшем рассмотрении их положение не подтверждается. Вопросом же, могут ли, а если да, то насколько, интервалы частот у разных (да и у одного и того же) автора «наезжать», или пересекаться друг с другом, супруги Фоменко как будто вообще не задавались. Если бы всё было действительно так, как они предписали, голос каждого писателя имел бы свою собственную частоту — то есть ими тогда был бы открыт замечательный инвариант идиостиля. По нашим данным, это не так: только у обсчитанных девяти авторов из их списка, в 38 произведениям, суммарные частоты пятерых значительно пересекаются: в интервале 21,43–22,64 % друг друга почти полностью перекрывают 3 текста Пушкина, 3 текста Набокова и 6 текстов Л. Толстого, с одной стороны; а с другой, в интервале чуть выше, 24,01–25,37 %, пересекаются 3 текста Гоголя и 4 текста Гончарова. (Напомню, что сама «процентная норма» берется целиком от всех слов в текстах соответствующих авторов: попросту сколько из них составляют служебные слова.) Это то, что можно было сказать по поводу магического списка из 54 служебных слов как *инварианте* авторского стиля. Но то же самое, как и следовало ожидать, происходит с отдельными элементами их списка. Так, оказываются очень похожими частоты союза А у семи из девяти авторов, а именно у двух пар и одной тройки: т. е. 1) Л. Толстого (47–85 %) и Набокова (47–84 %), 2) Пушкина (75–91 %) и Гоголя (55–99 %) и 3) Тургенева (76–142 %), Достоевского (96–136 %) и Гончарова (107–154 %). Сходная картина складывается по частице ТОЛЬКО: лишь интервалы частот у двоих, Пушкина (39–52 %)[9] и Горького (56–79 %) стоят обособленно (и ниже всех других), а вот частоты семерых остальных активно между собой пересекаются. Так, частота у Набокова (82–94 %) «наезжает» — на Тургенева (85–121 %), у Тургенева, в свою очередь — на Лескова (94–183 %), соприкасаясь с частотой Гончарова (116–141 %), причем частоты двоих последних пересекаются еще с Достоевским (131–144 %) и Л. Толстым (101–171 %). Наконец, на последнего из этой девятки, Гоголя (141–152 %) накладываются интервалы обоих, Толстого и Достоевского…

Итак, согласно данным ФиФ получалось, что при проценте употребления служебных слов у «несомненного Ш» в **22,5–24,4** %, этот же процент у «сомнительного Ш» (т. е. в первой половине ТД) — только **19,6** %! Таким образом, **три** или даже **пять** процентов расхождения кардинально отличали одного

---

[8] Согласно нашим данным, суммарная частота всех служебных слов по списку ФиФ в романах Достоевского (по трем текстам) действительно колеблется в очень узком интервале, даже еще более узком, чем полученный ими, но при этом — на 1% выше: «Идиот» — 26,42 %, «Преступление и наказание» — 26,56 %, «Братья Карамазовы» — 26,57 %.

[9] У Пушкина из всех самый малый корпус прозаических текстов — по 3 текстам около 70,3 тыс. слов.

от другого, а соответствующий параметр у К — **21**%, был подозрительно близок к показателю первой половины ТД: расхождение с частотой которого, как мы видим, составляло, по их данным, менее полутора процентов.[10] К тому же авторы предложили такой критерий:

> «Если для двух исследуемых произведений значения параметра 3 (процент служебных слов) разнятся больше, чем на единицу, то есть основания заподозрить различное авторство сравниваемых текстов».

(1,4% — все-таки несколько превышает единицу, не говоря уж о 3% или даже 5%.) Предложенная методика, или *критерий* ФиФ в данном случае как будто себя оправдывает, и все-таки, как мы убедимся ниже, не совсем.

§3. Определения

Тут пора ввести некоторые термины и дать определения. Для краткости все обсчитываемые ниже в таблицах единицы — служебные слова, предлоги, частицы, союзы, а также добавленные нами в список коннекторы [13], модальные, дискурсивные выражения и вводные конструкции именуются попросту **скрепами**.[11] Их частоты сравниваются с некими «нормативными» частотами, т. е., например, с уровнем частоты данной лексемы в ТД, только в его первой половине или же с частотой во всем Национальном корпусе русского языка (далее — НК)[12]. Помимо процентов удобно исчислять частоты скреп с помощью *промилле*, т. е. тысячных, а также миллионных долей (*миллипромилле*, ipm.: instance per million, или *промиллион*): в нашей таблице ниже частоты кроме абсолютных значений указываются или в ipm, или в процентах. **Отклонением** частоты f скрепы A у писателя X в тексте Y относительно писателя Z и его текста W будем называть простую разность частот fA(xy) — fA(zw). Такое отклонение естественно может быть как положительным, так и отрицательным. Отклонение fA(xy) от средней частоты по НК есть разность его с уровнем соответствующей частоты (fA(xy) — fср.НК). Нас более всего далее будет интересовать отклонение кого-то из фигурантов, Ш и К, от уровня частоты в первой части ТД (ТД1), т. е. разность между f(Ш)-f(ТД1) и f(К)-f(ТД1). Такие отклонения указаны в таблице ниже в процентах.

---

[10] Той же проблемой занимался еще и однофамилец супругов Фоменко — зав. кафедрой теории литературы Тверского университета (1994–2006) Игорь Владимирович Фоменко (1937–2015). Его работы на указанную тему [11, 12].

[11] Такое понятие, как «текстовая скрепа» — отчасти уже традиционно [14]; нами здесь оно трактуется расширительно.

[12] Там, где говорится о *норме НК*, на самом деле всюду имеется в виду просто уровень среднестатистического употребления лексемы, получаемый при запросе по адресу http://ruscorpora.ru/search-main.html. Следует учитывать, что основные расчеты частот в НК были нами произведены до середины февраля 2017, после чего корпус был пополнен примерно на 15 млн слов и сейчас составляет 283 431 966.

Например, союз КРОМЕ ТОГО более частотен в иных текстах, не художественных: если в НК его частота — 121,42 ipm., то в ТД1 — только 4,63, составляя всего лишь около 3% от уровня частоты по НК (см. Табл.); впрочем, и у Ш частота только чуть выше — 4,28%, зато у К — уже существенно выше: 19,10%. По этой частной скрепе их отклонения от ТД1 серьезно различаются, отклонение Ш почти совпадает с ней, а у К оно сильно отстоит: соответственно, на Ш+0,47 и К+15,29 (если же выражать отклонения в процентах, то: Ш=112% и К=501%), что в данном случае говорит в пользу авторства Ш. Всего же таких скреп, по которым предполагается измерять отклонения от ТД1, в нашем списке сейчас около тысячи. Выборочно четыре из них представлены на Таблице:

Скрепа / абс. и отн. частота (ipm.) по НК (100%)	**Частоты (абс. и отн.) в ТД1**	**в ТД2**	**у Ш**	**у К**	**Отклонения Ш и К (от ТД1) и их разность (Δ)**
А 2252989 8008,87	1682 7785,6 98%	2384 11542,89 146%	6002 15590,3 197%	6778 9825,01 124%	Ш+99 К+26 Δ = 73
…					
ВДРУГ 13800 519,97	40 185,15 36%	62 300,19 58%	154 400,02 77%	373 540,68 104%	Ш+41 К+68 Δ = 27
…					
КАК 1877428 6671,44	1391 6438,62 97%	1597 7732,38 116%	3361 8730,26 131%	4672 6772,27 102%	Ш+ 34 К+5 Δ = 29
…					
КРОМЕ ТОГО 32225 121,42	1 4,63 3,81%	1 4,84 3,99%	2 5,20 4,28%	16 23,19 19,10%	Ш+0,47 К+15,29 Δ = 14,82
…					

§3.1. Проверка гипотезы

Мы попробовали повторить подсчеты, воспроизведя и сам «эксперимент» супругов Фоменко, но только уже на полном материале текстов К, вернее, на всем доступном современному читателю (а полный К и сегодня не опубликован, так как многое ждет расшифровки в газетах столетней давности и кое-что еще покоится в архиве). Результаты эксперимента можно разделить на две части.

Во-первых, (оставив пока в стороне вопрос разделения текста ТД на две части) суммарные частоты служебных слов самих Ш и К в самом деле различаются между собой более, чем на 1%, и суммарная частота всех скреп по К оказывается все же **ближе** к суммарной частоте их же по ТД, чем та же величина у Ш. Разность суммарных частот по всем рассматривавшимся скрепам

для текстов К и ТД, т. е. Summa fК – Summa fТД = 21,96 % – 20,93 % = 1,03 %, в то время как для текстов Ш соответствующий показатель более чем вдвое выше: Summa fШ – Summa fТД = 23,36 % – 20,93 % = 2,43 %. Иначе говоря, *вывод*-1 из гипотезы ФиФ частично подтверждается.

Во-вторых, после того как мы все-таки разделили текст ТД на «сомнительного Ш» (ТД1) и «безусловно Ш» (ТД2), разница суммарной частоты во всех остальных текстах Ш и во второй половине ТД, т. е. f(Ш) – f(ТД2), составила всего лишь 1 %, тогда как разница того же Ш с первой половиной гораздо выше: f(Ш) – f(ТД1) = 24,29 – 20,75 = 3,54 %. Это вроде бы укладывается в версию о плагиате или же об использовании части рукописи К при написании ТД.

Но остается необъяснимым следующий результат: при том что суммарная частота служебных слов К расходится с первой половиной ТД более чем на 2 %: f(К) – f(ТД1) = (22,89 – 20,75) = 2,14 %, — что еще можно было бы объяснить искажением его стиля или же активной работой «соавтора», однако со второй половиной романа (в написании которого К никак не мог принимать участие) суммарная частота К расходится почему-то гораздо меньше — всего лишь на десятые доли процента: f(К) – f(ТД2) = 23,25 – 22,89 = 0,36 %! — причем отклонение оказывается меньше, чем отклонение самого Ш от ТД2! Можно ли объяснить это просто тем, что и данные ФиФ, и наши расчеты лежат в пределах статистической погрешности? Во всяком случае, остается фиксировать это как парадокс — для разрешения в дальнейшем.

§4. Расширение эксперимента и возможные предположения

Далее нами проделан следующий эксперимент: пытаясь применить к действительности гипотезу ФиФ о том, что наиболее различительными для идиостиля выступают как раз наименее значимые, наименее заметные составляющие языка, а именно частоты употребления союзов, предлогов и частиц, в список 54 нами были добавлены **вводные конструкции, междометия, модальные и дискурсивные слова**, а также **разрывные многосоставные коннекторы**, состоящие из нескольких слов, между которыми могут вклиниваться произвольные слова текста, от одного до 10 (например, такие как — *так* (…) *как*). В результате получилась таблица с частотами соответствующих лексем в НК, ТД, Ш и К, четыре строки из которой были приведены выше. Здесь понадобилось преодолеть следующее предубеждение: что частоты малозначимых, структурных элементов языка для всех говорящих на этом языке приблизительно одинаковы. Оказалось, что это не так, и попутно родились некоторые предположения, отчасти развивающие, но отчасти опровергающие то, что было названо нами выше *гипотезой* и *выводами* ФиФ, а именно:

(а) что по многим показателям один отдельно взятый писатель может быть надежно **отличён** от другого, если использовать частоты употребления им служебных слов (в более общей формулировке это та же гипотеза ФиФ);

(б) что один писатель всегда **более похож** на другого, нежели некий третий, т. е. что вполне можно сравнивать — даже количественно — писательские стили (но при этом являются ли все они так дискретно-комфортно

располагающимися на числовой шкале, как получалось в *выводах* 2 и 3 из гипотезы ФиФ, серьезные сомнения остаются: скорее всего это неверно);

(в) что частотные показатели могут быть **несимметричны**: т. е. если, например, какой-нибудь союз, например, *потому что* по относительной частоте у писателя А встречается вдвое чаще, чем у писателя Б, то суммарная частота его условных синонимов (т. е. в данном случае, группы союзов *так как, поскольку, ведь, ибо* итп.) вовсе не обязательно должна быть пропорционально выше у писателя Б;

(г) и более того: что для любого автора можно составить его **идиолектный профиль**, зафиксировав количественно отклонения — по тем или иным пунктам нашей таблицы, например, исчисляя *отклонение* от уровня частоты в НК:[13] но при этом гораздо удобнее оперировать не совокупным процентом сразу всех, как было у ФиФ, служебных слов, а — их расширенным списком, определяя отклонение частот каждой из «скреп» в отдельности, относительно какого-то избираемого ориентира; скажем, какой-то автор А надежнее отличается от Б не суммарной частотой служебных слов (fсум.А – fсум.Б) или же их процентом относительно всего текста, а — конкретным набором отклонений по таким-то и таким-то скрепам: потенциально каждой или хотя бы только одной из скреп списка (fa, fb, fc,...), что и составляет в целом его **идиолектный профиль**, или, если угодно, *портрет*, фиксируя нам как бы «отпечатки пальцев» данного автора или его произведения.

К примеру, известное авторское словечко Достоевского **ВДРУГ**, подсчитанное по трем его романам, может быть следующим образом отражено в его профиле: употребляясь в «Идиоте» с частотой 3042 ipm., в «Преступлении и наказании» — 3325, а в «Братьях Карамазовых» — 3969. Сравнивая эти частоты с Л. Толстым, видим, что для последнего само слово, напротив, крайне нехарактерно: в «Войне и мире» — 1114, «Анне Карениной» — 989, а в «Воскресении» — вообще 631 (при том, что средняя частота по всему НК — еще меньше: 520, т. е. ниже даже Толстовской по Воскресенью)...

Перечисленные выше пункты (а–г), являются прогностическими конструктами — возникшими на основе наблюдаемых фактов: их можно считать уже нашими собственными гипотезами, или *предположениями*, которые рождались в ходе работы и нуждаются в подтверждении/опровержении при продолжении эксперимента.

§5. Итог с подсчетом по методу ФиФ

Если итоговые частоты из ipm. перевести в проценты, получается, что суммарный уровень частот всех 54 скреп по НК — 22,22%; а суммарная частота

---

[13] Весь НК берется нами просто как удобная точка отсчета, или «репер». Возможно избрать другие — частоты в художественных текстах или в текстах тех авторов, «кругу» которых принадлежал данный...

их в ТД, без разделения романа на части, — 20,93 %; в 1-й половине ТД — 19,75 %; а во 2-й — 22,16 %; у Ш — 23,36 %, а у К — 21,96 %. То есть ближе всего ТД2 оказывается К: с разницей в десятые доли процента, а вот с ТД1 у него расхождение более чем на 2 % (21,96 – 19,75 = 2,21), тогда как у Ш это расхождение еще больше (23,36 – 19,75 = 3,61). Сопоставив результаты с полученными исследователями в прошлом веке, можно сказать, что и с привлечением новых данных их *вывод*-1 пока остается верен, но разрыв между К и Ш заметно сократился и расхождение между К и ТД1 перестало удовлетворять *критерию* ФиФ (см. §2).

§5.1. Итог с подсчетом по отклонениям

21 скрепа из заданных в списке 54 единиц указывает в пользу авторства Ш (итоговое отклонение в этих скрепах у Ш меньше, чем у К), а 30 — в пользу авторства К (в них у Ш отклонение больше — как у скреп А и КАК в Таблице выше — в отличие от скреп ВДРУГ и КРОМЕ ТОГО, для которых больше отклонение у К); при этом три скрепы — как-бы «ничьи» (в них Ш≈К, или расхождение не превышает 1 %). Получается, что, как и утверждали ФиФ, по сумме отклонений идиолект Ш отстоит от идиолекта ТД все-таки дальше, чем К, но, правда, это расхождение весьма незначительно, составляя всего 2,2 % от общей суммы отклонений. Можно считать, что окончательный ответ, является ли *вывод*-1 правильным, здесь не опровергнут, но и не подтвержден.

§6. Особое мнение С. Л. Николаева

Примененная Михеевым и Эрлихом процедура разграничения авторских идиолектов, восходящая к ФиФ, привела к отчасти абсурдным выводам: большая часть служебных слов имеет приблизительно одинаковую частоту в сравниваемых корпусах; в оставшемся материале значительное число слов имеет частоты, которые свидетельствует о том, что автором ТД2 мог быть Крюков, что невероятно. Мне кажется, что эта процедура не может привести к достоверным выводам по той простой причине, что большинство служебных слов выполняет грамматическую функцию и обязательны для употребления в тексте на естественном языке. Разница авторских идиолектов по частоте использования служебных слов связана в первую очередь с разной пропорцией синонимов (другие причины различия — пристрастие к простым/сложным предложениям, от чего зависит частота союзов и ряда частиц).

Синонимы — лексемы (слова и устойчивые словосочетания), имеющие похожее лексическое значение. Каждый синоним имеет свой особый оттенок значения, отличающий его от других синонимов, например: *красный — алый — багряный — багровый*. Однако благодаря сходству значений синонимы зачастую используются для передачи одинакового смысла. Особенно это относится к служебным лексемам-синонимам: *так как — потому что, по причине — вследствие, некто — кто-то, только — лишь* и т. д. Богатство синонимии в языке отражает тенденцию к языковой избыточности. Однако в естественном языке одновременно действует тенденция к экономии языковых

средств. Поэтому в каждом конкретном языке, диалекте, идиолекте одни синонимы используются реже других. Например, одни люди чаще употребляют *только, так как, по причине*, а другие — их синонимы *лишь, потому что, вследствие* и т. п. Система предпочтений синонимов строго индивидуальна для каждого идиолекта, и при достаточной выборке материала система количественных отношений между синонимами как правило демонстрирует инвариантный «идиолектный профиль» человека, который пропорцией хотя бы в одной из синонимических пар отличается от любого другого носителя данного языка (диалекта). Эта модель подходит и для различения «авторских идиолектов» — т. е. языков, отраженных в художественных текстах. Что касается конкретных корпусов ТД1, ТД2, Ш и К, то между ними отмечаются следующие закономерности пропорций между синонимическими «гнездами», выбранными из материала Михеева и Эрлиха.

I. Синонимические лексемы, пропорция которых равна во всех корпусах.

II. Во всех остальных примерах пропорции в ТД2 и у Шолохова приблизительно одинаковые, что доказывает идентичность идиолектов ТД2 и Шолохова — иными словами, 2-я половина «Тихого Дона» написана Шолоховым.

IIа. Как правило, тексты Крюкова кардинально отличаются от текстов ТД2/Шолохов по пропорциям синонимов. Также чаще всего пропорции синонимов в ТД1 отличаются как от ТД2/Шолохов, так и от Крюкова, что, по всей видимости, говорит о том, что автором ТД1 не были ни Шолохов, ни Крюков.

IIб. Сюда примыкают еще 2 типа комбинаций, в которых разные идиолекты, по-видимому, случайно совпадают попарно по пропорциям синонимов:

1. Примеры, в которых пропорции Крюкова близки к пропорциям ТД/Шолохов и одновременно к ТД1.
2. Примеры, в которых пропорции Крюкова близки к пропорциям ТД1, но отличаются от пропорций ТД2/Шолохов.
3. Примеры, в которых пропорции в ТД1 близки к пропорциям в ТД2/Шолохов, но резко отличаются от пропорций Крюкова.

Согласно статистике частот так называемых «скреп» и в особенности по пропорциям синонимов Крюков является весьма сомнительным автором ТД1 и не имеет лингвистического отношения к ТД2. Однако в качестве автора ТД1 не годится и Шолохов: он тоже с трудом проходит и по «скрепам», и по синонимическим пропорциям. Однако в общем и целом ТД1 и ТД2/Ш ближе между собою и вместе противопоставлены Крюкову. Разумным решением этой контроверзы может быть предположение о том, что Шолохов обработал чужой (но не крюковский!) текст, известный нам как ТД1, а ТД2 написал в стиле ТД1. Что касается текстуальных параллелей между ТД и текстами Крюкова, то они встречаются в основном в ТД1. Поэтому не исключено, что именно «автор ТД1» был знаком с текстами Крюкова. «Творящий в застенках ОГПУ писатель» — всего

лишь красивая легенда, но гипотеза, что первая половина ТД является текстом неизвестного нам талантливого автора «под редакцией» Шолохова, может оказаться не столь уж экстравагантной. Ею могут объясняться как выдающиеся художественные достоинства ТД1 (не характерные ни для тексов Крюкова, ни тем более Шолохова), так и особенности языка.

§7. Заключение

Итак, мы вынуждены признать, что на ограниченном материале, который обсчитывался ФиФ, по сравнению только суммарной их частоты, как они сами это делали, а также с учетом отклонений частот каждой из скреп, а кроме того еще и просто общим сравнением количества скреп в пользу того и другого из претендентов их *вывод*-1, с некоторой натугой, но — проходит (§5 и §5.1). Однако это не означает, во-первых, того, что и на расширенном материале он также окажется верен — хотя бы на той тысяче служебных слов и их сочетаний, которую мы предлагаем в качестве анкеты для «идиолектного профиля». Во-вторых, он может быть опровергнут увеличением числа рассматриваемых авторов, иначе говоря, нет полной уверенности, что **между** Ш и К не «втешется» еще кто-нибудь третий, чей «профиль» окажется ближе к ТД, нежели профили обоих обсуждаемых в течение последнего почти что столетия претендентов.

Для объективной оценки близости к идиостилистике ТД должны быть установлены, во-первых, по возможности полные корпуса авторов, во-вторых, привлечено значительно большее число языковых скреп, чем это было у ФиФ, и в-третьих, сами скрепы должны быть оценены с точки зрения их различительности, например, будучи выстроены по убыванию разницы между f(К) – f(ТД1) и f(Ш) – f(ТД1) и между f(К) – f(ТД2) и f(Ш) – f(ТД2). В этом случае будет удобно сравнивать общие количества скреп с отклонениями и величины общих сумм таких отклонений. Судить о возможности написания ТД1 каким-то третьим лицом, помимо Ш и К, до выяснения вопроса об использовании/не-использовании при его написании некого «пред-текста» К, на наш взгляд, преждевременно.

* * *

При написании данной статьи ее авторам оказывал неоценимую помощь — как в том, что касалось критики, ценных советов, оттачивания формулировок, так и в придании ей научного вида в соответствии с требованиями современной лингвостатистической теории — **Сергей Львович Николаев**, которому мы искренне благодарны. Выражаем признательность также **Борису Валерьевичу Орехову** — за указание полезной по теме литературы и **Дмитрию Владимировичу Сичинаве** — за подсказку далеко не простой возможности подсчета в НК слов с дефисом.

Литература

1. *Ф. Д. Крюков.* Над обрывом. Очерки и статьи последних лет жизни (1917–1919). М.—СПб. 2009 — http://uni-persona.srcc.msu.su/f-krukov/.
2. «Рабочая газета» 24 марта 1929 — из письма Серафимовича, Авербаха, Киршона, Фадеева и Ставского в защиту Шолохова (перепечатано «Правдой» 29 марта 1929).
3. *Андрей Чернов.* КАК СПЕРЛИ ВОРОВАННЫЙ ВОЗДУХ. Заметки о «Тихом Доне» https://nestoriana.wordpress.com/2016/02/03/vorovanyi-vozdyh/
4. *А. В. Венков,* «Тихий Дон»: источниковая база и проблема авторства. Ростов-на-Дону, 2000.
5. *«Д».* Стремя «Тихого Дона». Загадки романа. — Paris: YMCA Press. 1974.
6. *Г. Хьетсо, С. Густавссон, Б. Бекман, С. Гил.* Кто написал «Тихий Дон»? М., 1989.
7. *Л. З. Аксенова (Сова), Е. В. Вертель.* О скандинавской версии авторства «Тихого Дона» // на сайте МГУ: http://www.philol.msu.ru/~lex/td/?pid=012116&oid=01211.
8. *А. Г. Макаров, А. А. Поликарпов*, «О неоднородности количественных характеристик...» (2010) http://www.philol.msu.ru/~rlc2010/abstracts/rlc2010_abstracts_sec14.pdf.
9. *З. Бар Селла*. Литературный котлован. Проект «Писатель Шолохов», М.:РГГУ, 2005. 462 с.
10. *Фоменко Т. Г. и В. П.* Авторский инвариант русских литературных текстов. Приложение. Кто был автором «Тихого Дона»? — http://lib.ru/FOMENKOAT/greece.txt.
11. *И. В. Фоменко.* Практическая поэтика. Москва : Академия, 2006. — 191 с.
12. *И. В. Фоменко, Л. П. Фоменко.* Художественный мир и мир, в котором живет автор // Литературный текст: проблемы и методы исследования. IV: сборник научных статей / под ред. Л. В. Тарасова. — Тверь, 1998.
13. *О. Ю. Инькова Н. А. Попкова.* Структура двухместных коннекторов русского языка в свете корпусных данных // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2016» Москва, 2016 — http://www.dialog-21.ru/media/3394/inkovaoyu.pdf.
14. *Прияткина, А. Ф.* Текстовые «скрепы» и «скрепы-фразы» (о расширении категории служебных единиц русского языка) // А. Ф. Прияткина. Русский синтаксис в грамматическом аспекте (синтаксические связи и конструкции). Избранные труды. — Владивосток: Изд-во Дальневост. ун-та, 2007. — С. 334–344.
15. *Burrows J.* Questions of authorship: attribution and beyond // Computers and the Humanities, 2003, 37, p. 1–26.