УДК 81'33

СТИЛЕМЕТРИЧЕСКОЕ ИССЛЕДОВАНИЕ ТЕКСТОВ УЧАСТНИКОВ ЭКСТРЕМИСТСКОГО ФОРУМА: ГЕНДЕРНЫЙ АСПЕКТ[1]

ЛИТВИНОВА Татьяна Александровна,
кандидат филолологических наук,
заведующая научно-исследовательской лабораторией корпусной идиолектологии,
Воронежский государственный педагогический университет

АННОТАЦИЯ. *Проблема исследования экстремистских текстов относится к одной из актуальных проблем современной российской лингвистики, однако фигуре авторов таких текстов, комплексному анализу их языковой продукции, в том числе в гендерном аспекте, уделяется гораздо меньше внимания. Кроме того, при анализе подобных текстов редко используется инструментарий современной корпусной и компьютерной лингвистики. В статье представлены результаты стилеметрических (т.е. основанных на анализе поддающихся квантификации элементов текста) экспериментов, направленных на выявление характерных особенностей текстов мужчин и женщин – участников форума, внесенного в Федеральный список экстремистских материалов РФ. Анализ постов форума, проведенный с использованием инструментов современной компьютерной лингвистики, позволил выявить как общие характеристики подобных текстов, так и гендерно специфичные, а также показал перспективность указанного подхода к анализу текстов интернет-коммуникации, содержащих противоправный контент.*
КЛЮЧЕВЫЕ СЛОВА: *корпус текстов, экстремистский текст, стилеметрия, корпусная лингвистика, гендерная лингвистика, классификация текстов.*

A STYLOMETRIC STUDY OF THE EXTREMIST FORUM POSTS: GENDER DIMENSION

LITVINOVA T. A.,
Cand. Philolog. Sci., Head of Corpus Idiolectology Lab,
Voronezh State Pedagogical University

ABSTRACT. *The problem of analysis of extremist texts is one of the urgent problems of modern Russian linguistics. Much less attention, however, is paid to the figure of the authors of such texts, namely comprehensive analysis of their language production, including gender dimension. In addition, efficient tools of modern corpus and computer linguistics are rarely used in analysis of such texts. The article presents the results of stylometric (i.e., based on the analysis of quantifiable text elements) experiments aimed at identifying the characteristic features of the texts of male and female authors of posts from forum included in the Federal List of Extremist Materials of the Russian Federation. An analysis of the forum posts with the tools of modern computer linguistics allowed us to identify both general and gender-specific characteristics of such texts, as well as to show the advantages of this approach to the analysis of Internet communication texts containing illegal content.*
KEY WORDS: *text corpus, extremist text, stylometry, corpus linguistics, gender linguistics, text classification.*

Введение

В связи с распространением интернет-коммуникации проблема исследования экстремистских текстов, т.е. текстов, содержащих «публичное оправдание терроризма либо публичные призывы,провоцирующие рознь и ненависть, призывы к насилию, или же направленных на возбуждение социальной, национальной или религиозной розни; пропаганду исключительности, превосходства либо неполноценности человека в зависимости от его принадлежности к определенному этносу, расе, от его вероисповедания, от его родного языка» (ФЗ от 25 июля 2002 г. N 114-ФЗ «О противодействии экстремистской деятельности»), приобрела особую актуальность. Исследователи разрабатывают различные методы анализа экстремистских текстов на русском языке (см., например, многочисленные публикации по данной проблеме в журнале «Политическая лингвистика», многочисленные материалы конференций и круглых столов[2] и т.д.), однако данная область, несмотря на большое число работ, характеризуется отсутствием синергии между «качественными» и «количественными» методологиями анализа текста. Большинство работ по указанной проблематике выполнено в русле «качественной» методологии, предполагающей «ручной», экспертный анализ текста, что, с одной стороны, по-

[1] Исследование выполнено при поддержке гранта Президента РФ № МК-5718.2018.6 «Речевой портрет экстремиста: корпусно-статистическое исследование (на материале экстремистского форума "Кавказчат")».

Информация для связи с автором: centr_rus_yaz@mail.ru

[2] Так, проблеме анализа экстремистских текстов были посвящены два секционных дня VI Международного конгресса исследователей русского языка "Русский язык: исторические судьбы и современность" (http://rlc2019.philol.msu.ru/). Данная секция была одной из самых многочисленных – как по числу докладов, так и по количеству участников.

зволяет выявить типологические особенности экстремистских текстов на трудно поддающихся формализации уровнях семантики и прагматики, но с другой – ограничивает объем проанализированного материала, а также в ряде случаев существенно снижает уровень воспроизводимости полученных результатов. Количественные методы анализа применяются к подобного рода текстам на русском языке в ограниченном числе работ (см. подробнее: [9]), хотя такой подход является достаточно распространенным в области анализа текстов экстремистского содержания на европейских языках (см. подробнее: [7]). Однако данный подход также не лишен недостатков: основное внимание в подобных работах уделяется созданию классификаторов, позволяющих относить текст к классам «экстремистский» vs. «неэкстремистский», однако мало внимания уделяется собственно лингвистическим признакам экстремистского дискурса (см. подробнее: [1]). На наш взгляд, сочетание качественной и количественной методологии анализа экстремистских текстов, в частности активное применение методов корпусной лингвистики, необходимо как для теоретического осмысления феномена экстремизма и его отдельных проявлений, так и для прикладных разработок в области автоматического распознавания экстремистского контента.

Экстремистский текст не существует отдельно от его автора. Более того, фигура автора текста в условиях перехода к «персонализированной» пропаганде экстремистской идеологии, ведущейся посредством интернет-коммуникации, становится особенно значимой: «…"я" актуализируется в современном экстремистском дискурсе» [3]. Однако проблема исследования речевого портрета автора текста экстремистского содержания является мало изученной [7], между тем как анализ речевой продукции (не только экстремистских текстов) таких лиц позволит значительно расширить существующие представления о личностных особенностях и ценностных установках указанной группы.

Отдельным малоисследованным аспектом проблемы автора экстремистских текстов является гендерный. Несмотря на то что в литературе утверждается положение о том, что экстремистский автор – это, как правило, лицо мужского пола (см., например, [2]), известно, что, например, количество женщин, поддерживающих деятельность Исламского государства (ИГ, запрещено в РФ), неуклонно растет [14]. Женщины могут быть не только объектом, но и субъектом пропаганды, особенно в условиях интернет-коммуникации, где важно установить доверительный контакт с собеседником [3]. Следовательно, указанное выше положение о половой принадлежности автора экстремистского текста нуждается в пересмотре.

Таким образом, проведенный нами краткий анализ литературы позволяет утверждать, что проблема поиска сходств и различий в текстах мужчин и женщин – авторов экстремистского форума на уровне квантифицируемых параметров текста с последующей содержательной интерпретацией результатов стилеметрического анализа является актуальной проблемой современного российской лингвистики.

1. Материал и методы исследования

1.1. Материал исследования

Материалом исследования являются посты участников форума «Кавказ-Чат», входящие в в состав датасета ‘kavkazchat’, являющегося частью обширной коллекции документов, использующейся исследователями для решения различных задач анализа текстов (подробнее о коллекции см. [6]). Форум «Кавказ-Чат» внесен в Федеральный список экстремистских материалов РФ. Ряд участников форума открыто выражают ненависть к России и ее гражданам. Исследователи относят данный форум к одним из инструментов «террористической агитации и пропаганды» [8].

Датасет ‘kavkazchat’ содержит 699981 постов, написанных 7125 участниками и входящих в одну из 16854 тем. Следует, однако, отметить, что наряду с темами, напрямую связанными с обсуждением «героев кавказского джихада», террористических актов и т.д., форум содержит темы, связанные с обсуждением кулинарии, семейной жизни и т.д. Далеко не все авторы форума открыто выражают экстремистские взгляды. Следовательно, необходимо выработать критерии отбора авторов для решения задач настоящего исследования. Кроме того, единые критерии отбора материала необходимы для корректного компаративного анализа. Для целей настоящей работы нами был сформирован корпус текстов, принадлежащий авторам, которые: 1) создали не менее 50 постов; 2) в постах которых не менее трех раз встречаются совместно леммы русн*+куфр*; и/или русн*+джихад*; джихад+к*фир*. Выбор данных лемм основывался как на нашем предварительном анализе материала, так и на работах исследователей, выполненных на мультиязычном материале, в которых было показано, что частотности лемм «kaffir» («неверующий»; в постах анализируемого форума встречаются написания кафир, куфр, куфир), jihad («борьба») являются одними из признаков, наиболее значимых для выявления твитов авторов, поддерживающих ИГИЛ (запрещена в РФ) [11]. Русня – характерное для данного форума пренебрежительное наименование России. Ниже (рис. 1) в качестве примеров приведем типичные контексты употребления данного слова авторами-женщинами, извлеченные при помощи пакета quanteda [10].

Авторы форума, отвечая на то или иное сообщение, копируют его; посты были очищены от таких повторяющихся сообщений. Тексты авторов, не соответствующих указанным выше критериям, были также удалены, после чего тексты отобранных авторов были просмотрены вручную (так, были исключены тексты участников форума, вступающие в спор с авторами, выражающими антироссийские взгляды, и, следовательно, использующие указанные выше слова в контекстах вида «вы называете нас кафирами»). Итоговый корпус содержит тексты 109 авторов, из которых 16 авторов было отнесено нами к авторам-женщинам (на основании анализа текстов, в которых автор открыто указывает свою половую принадлежность: «я как девушка…», «мы, женщины…», а также на основании информации о пользователях, содержащейся в постах других авторов). Для сравнительного анализа нами было отобрано 16 авторов-мужчин, при этом при выборе авторов-мужчин мы ориентировались на объем текста, так что каждому автору-женщине соответствовал автор-мужчина с примерно равным объемом текста (табл. 1).

Все тексты пользователя объединялись в один файл. Таким образом, наш подход является не тексто-, а авторо-ориентированным, то есть направленным на анализ всего массива текстов, созданных отобранными авторами экстремистского форума.

```
> print(kwic(corpus1, window = 4, valuetype = "glob", separator = " ", case_insensitive = TRUE, pattern = "русн*"))

      [elena_krim_F.txt, 781]                    вы это понимаете. |        Русня         | сегодня отчиталась, что
     [elena_krim_F.txt, 5409]                        , как и в |        русне         | - попробуй избавиться теперь
     [elena_krim_F.txt, 6475]                  уже нагнали со всей |        русни         | . Надо или все
     [elena_krim_F.txt, 8589]                           чего. Ну, |        русня         | ! Нарвется однажды,
    [elena_krim_F.txt, 10000]                     завалами. Траур в |        русне         | уже не перманентный,
    [elena_krim_F.txt, 10051]           и вся внутренняя политика |        русни         | . Молари теперь ХУНАФОЙ
    [elena_krim_F.txt, 11137]                        , а еще и |        русне         | надо угодить. Попал
    [elena_krim_F.txt, 13098]                  ЧЕМОДАН, ВОКЗАЛ, |        РУСНЯ         | - я таким говорю
    [elena_krim_F.txt, 14481]              бысро восстановят. Это |        русня         | так и лежала бы
    [elena_krim_F.txt, 18598]                  вы все про развал |        русни         | ! Тут Украина изчезает
    [elena_krim_F.txt, 21014]          необходимо рвать связи с |       русней         | .- Хотя бы
    [elena_krim_F.txt, 22579]             так Саакашвили хоть с |       русней         | не водится. А
    [elena_krim_F.txt, 27674]                  , кто враг давно |        русне         | , Тех, у
    [elena_krim_F.txt, 32798]                  такая реклама. И |        русня         | с радостью приютила такого
    [elena_krim_F.txt, 34867]               бы такими агентами у |        русни         | , то страны бы
    [elena_krim_F.txt, 35375]              то это сделали власти |        русни         | , чтобы свалить на
        [Muslima_F.txt, 1294]                        и они же до |        русни         | ещё первыми приняли удар
        [Muslima_F.txt, 7763]           обрусевшим, для которых |        русня         | только белая и пушистая
        [Muslima_F.txt, 7892]              машааЛЛАХ1) Так скоро |        русня         | превратится в Имарат Русь
      [Muslima_F.txt, 14892]             очень хорошо научились у |      русняцких       | чекистов- так и
      [Muslima_F.txt, 22111]                          ! да и в |        русне         | такие есть..
 [rabynya_allaha_F.txt, 2768]          , позорная, подлизывавшая |        русне         | ) лично для меня
          [Sura_F.txt, 3149]              духовном плане община в |        Русне         | . Новое поколение мусульман
          [Sura_F.txt, 3219]        и моральное разложение кафиров |        Русни         | пугают таких как Путины
          [Sura_F.txt, 3342]     не обольщаемся гнилым состоянием |        Русни         | . Дело не в
          [Sura_F.txt, 6465]                 не воюют конкретно с |       русней         | ... Наверное
          [Sura_F.txt, 8303]                    , но выступет за |        русню         | , если я не
          [Sura_F.txt, 8332]                   то не хочется за |        русню         | болеть...
```

Рис. 1 – Типичные контексты употребления слова «Русня»

Таблица 1 – Сравнительный анализ

Число авторов	Общий объем текста (в токенах)	Средний объем документа (в токенах)	Минимальный объем документа (в токенах)	Максимальный объем документа (в токенах)
16 мужчин	332 310	26651 (SD = 39892.22)	1397	160972
16 женщин	356 339	27315 (SD = 41454.11)	2328	171041

1.2. Методы исследования

Нами были использованы современные методы автоматической обработки текста. Одним из важных этапов автоматического анализа текстов является лемматизация, то есть приведение словоформы к лемме, т.е. словарной форме. Однако вследствие того что анализируемые тексты содержат большое число иноязычных слов, написанных на кириллице, ошибок, сленга и т.д., что значимо ухудшает качество лемматизации, было принято решение провести две серии экспериментов: с лемматизацией и без лемматизации. Лемматизация проводилась с использованием программы mystem[1]. Дальнейший анализ текстов проводился в среде R с использованием преимущественно двух пакетов, специально предназначенных для анализа текстов, – quanteda [10] и Stylo [12].

2. Результаты экспериментов и обсуждение

2.1. Эксперименты на корпусе с лемматизацией

Эксперименты на корпусе с лемматизацией проводились с использованием пакета quanteda. Стоп-слова[2] были удалены с использованием словаря стоп-слов, входящего в состав пакета (отметим, что он содержит не леммы, а словоформы, причем далеко не исчерпывающий их список; так, в нем содержатся словоформы эти, эту, но отсутствует словоформа это). На рис. 2 представлено облако слов, на котором изображены 100 самых частотных лемм корпуса, содержащиеся не менее чем в 10% текстов (чем крупнее шрифт, которым изображена лемма на облаке слов, тем более частотной она является) и дающие представление об основных темах корпуса (без разделения на тексты мужчин и женщин).

Перейдем к анализу текстов в гендерном аспекте. На рис. 3 представлены 15 самых частотных лемм в текстах мужчин и женщин (абсолютная частота; ранг; число документов, в которых встречается лемма). Данные наглядно показывают, что качественно списки самых частотных лемм в постах женщин и мужчин практически совпадают, за некоторым исключением. В топ-15 самых частотных лемм «женского» подкорпуса входит наречие-интенсификатор очень, а также лемма мусульманин, в то время как для мужских текстов характерно наличие большего числа глаголов в списке самых частотных лемм, а также лемма брат, которая часто используется в постах форума для обращения к собеседнику.

Если же мы посмотрим на дальнейшие уровни частотного списка, то обнаружим некоторые гендерные различия в лексическом выборе (рис. 4). «Компаративное» облако слов основано на сравнении частот слов в двух подкорпусах, при этом чем крупнее шрифт, тем больше разница между частотами слов. Как мы видим, различия в абсолютных частотах слов показывают, что в текстах мужчин более частотны леммы кафир, шахид, джихад, а также связанные с «русской» темой, органами власти и правопорядка, войны; в текстах женщин –

[1] https://yandex.ru/dev/mystem/ (дата обращения: 11.11.2019)

[2] В основном служебные части речи и местоимения, наречия-интенсификаторы, оценочные прилагательные.

леммы, связанные с обозначением чувств, «семейной» темой.

Однако разница в абсолютных частотах не всегда является информативной, в связи с чем мы воспользовались функцией пакета quanteda textstat_keyness, направленной на поиск ключевых слов, основанный на статистических тестах.

Рис. 2 – 100 самых частотных слов корпуса

	feature	frequency	rank	docfreq	group
1	это	2571	1	16	F
2	аллах	1506	2	15	F
3	который	1447	3	16	F
4	свой	1430	4	16	F
5	весь	1386	5	16	F
6	мочь	975	6	16	F
7	время	895	7	16	F
8	наш	754	8	16	F
9	год	723	9	15	F
10	знать	715	10	16	F
11	очень	712	11	15	F
12	тема	668	12	14	F
13	говорить	623	13	16	F
14	день	597	14	16	F
15	мусульманин	582	15	14	F

16	это	2388	1	16	M
17	который	1928	2	16	M
18	аллах	1879	3	16	M
19	свой	1493	4	16	M
20	весь	1373	5	16	M
21	год	923	6	16	M
22	мочь	888	7	16	M
23	время	776	8	16	M
24	говорить	748	9	16	M
25	тема	703	10	15	M
26	наш	696	11	14	M
27	давать	683	12	16	M
28	становиться	666	13	16	M
29	знать	608	14	16	M
30	брат	602	15	15	M

Рис. 3 – Самые частотные слова в текстах женщин (F) и мужчин (M)

Рис. 4 – «Компаративные» облака слов:
а – с максимумом в 50 лемм, б – с максимумом в 200 лемм;
F – слова, характерные для постов авторов-женщин;
M – слова, характерные для постов авторов-мужчин

Ключевые слова выделяются нами на основании сопоставления частот лемм в «мужском» и «женском» подкорпусах с использованием статистических критериев. На рис. 5 приведены ключевые слова, отобранные по критерию хи-квадрат с поправкой Йетса, $p < 0{,}001$ (по мере «приближения» к центру значение критерия убывает).

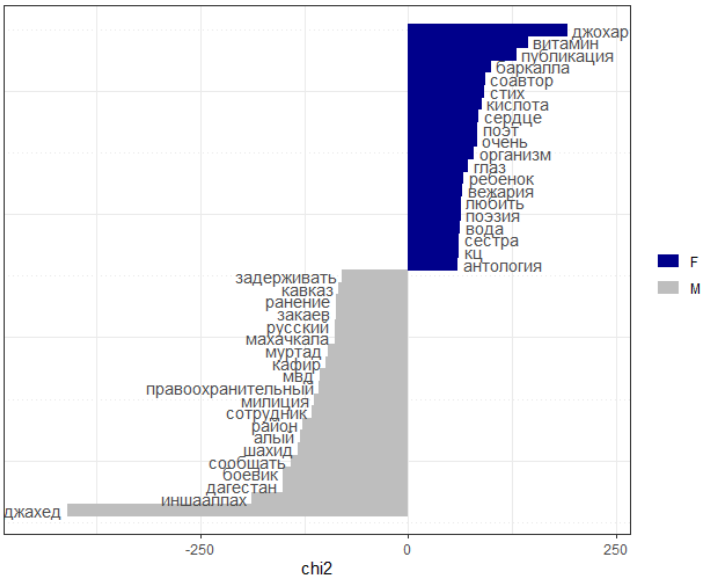


Рис. 5 – Ключевые слова «мужских» и «женских» текстов

Ряд наблюдений, сделанных при помощи инструмента «облако слов», согласуется с результатами анализа, направленного на выделение ключевых слов. Так, статистически значимыми оказываются различия в использовании слов, тематически связанных с Россией («русский»), органами правопорядка (задерживать, милиция, мвд) и темой войны (шахид, боевик, ранение, муджахед) (эти слова более частотны в текстах мужчин); слов, связанных с литературой (стих, поэзия, поэт, антология, соавтор), чувствами и «телесной» темой (любить, сердце, организм, глаз), темой семьи (ребенок) (более частотны в текстах авторов-женщин). Авторы-мужчины чаще используют слова, обозначающие неверующих (кафир) либо вероотступников (муртад), а также восклицание иншаллах. Как мужчины, так и женщины используют географические наименования (Дагестан, Кавказ, Махачкала – мужчины, Вежария (село в Дагестане) – женщины); женщины также используют чаще имя собственное Джохар (речь идет о Джохаре Дудаеве, чеченском политическом деятеле).

Приведенные выше наблюдения касаются частот отдельных слов. Дальнейшим этапом нашего исследования являлся анализ, направленный на выявление статистически значимых различий между текстами мужчин и женщин по параметрам, отражающим частотности в тексте слов определенных групп. В частности, нами ранее были составлены и апробированы словники для программы LIWC (см. подробнее: [5]), такие как Ego (в него вошли слова, обозначающие автора: я, по-моему и т.д.); словарь дискурсивных маркеров; словарь интенсификаторов (очень, сильно), дейктиков (там). Кроме того, нами был рассчитан параметр «число слов длиннее 6 букв». Указанные группы слов были выбраны на основании работ, в которых анализируются различия между авторами экстремистских текстов и контрольными текстами (см., например, работу [15]), а также литературы, связанной с поиском различий в текстах женщин и мужчин (см. обзор в монографии [4]).

Результаты расчета с использованием критерия Уилкоксона для независимых выборок, выполненных в R, приведены в табл. 2.

Таблица 2 – Результаты расчета с использованием критерия Уилкоксона

Параметр	Категория	Значение критерия	Наличие различий по параметру в текстах мужчин и женщин
Слова длиннее 6 букв	Сложность	W = 83, p = 0.09351	-
Дискурсивные маркеры со значением причины	Сложность	W = 74, p = 0.04342	+
Интенсификаторы	Эмотивность	W = 176, p = 0.07292	-
Дейктики	Пространственный дейксис	W = 140, p = 0.6647	-
Эго	Я	W = 212, p = 0.001648	+

Таким образом, нами были выявлены различия в частотах слов категории «Я» и дискурсивных маркеров со значением причины (см. также рис. 6).

Более частое использование слов категории «Я» женщинами неоднократно отмечалось исследователями, работающими с текстами на разных языках [4]. В целом различий в использовании интенсифи-

каторов нами отмечено не было, несмотря на наличие статистически значимых различий в использовании одного из слов этой категории (очень), что показывает необходимость анализа как отдельных слов, так и категорий в целом при описании речевого портрета группы лиц.

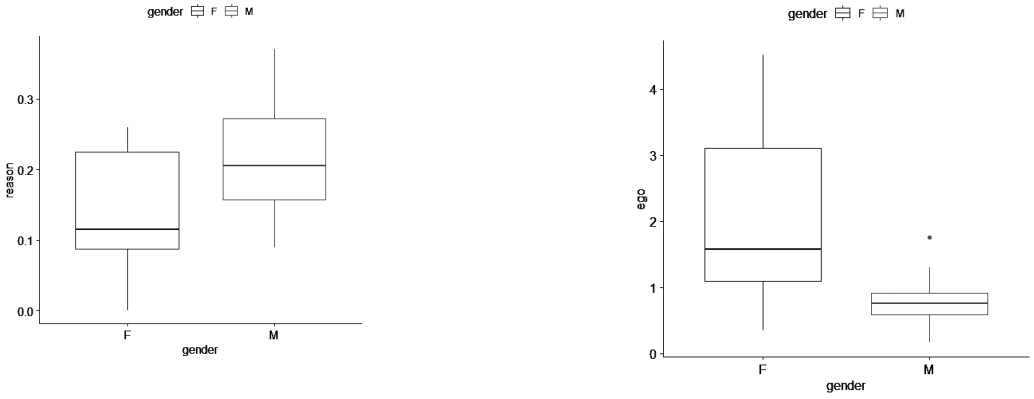
Что касается категории сложности, то в некоторых работах отмечается, что когнитивная сложность в текстах экстремистских авторов является более высокой в сравнении с текстами условной нормы [15]. В нашем материале отмечается более высокий уровень встречаемости дискурсивных маркеров со значением причины в текстах авторов-мужчин, однако не обнаружилось различий по параметру «слова длиннее 6 букв». Как нам представляется, дальнейшие исследования уровня когнитивной сложности необходимы как в гендерном аспекте, так и в аспекте исследования экстремистских текстов.

Таким образом, проведенный нами анализ позволил выявить как сходства, так и различия на лексическом уровне текстов мужчин и женщин – участников экстремистского форума. Анализ проводился нами на уровне лемм. На следующем этапе мы проводили эксперименты на корпусе без лемматизации.

2.2. Результаты экспериментов на корпусе без лемматизации

Целью экспериментов на корпусе без лемматизации являлась как проверка результатов, найденных на предыдущем этапе, так и оценка значимости различий между текстами женщин и мужчин путем использования различных методов, в том числе техник классификации текста. Все эксперименты с корпусом без лемматизации проводились с использованием пакета Stylo [12].

На первом этапе нами проводился кластерный анализ текстов мужчин и женщин. Кластерный анализ представляет собой одну из техник, направленных на поиск структуры в данных. На данном этапе мы использовали в качестве параметров текста стандартизованные частоты 100 самых частотных слов корпуса, в качестве меры расстояния – различные метрики, реализованные в указанном пакете. На рис. 7 представлена дендрограмма, построенная с использованием в качестве меры расстояния дельты Эдера, которая рекомендуется для флективных языков [12]. Данная мера является модификацией стандартной меры Берроуза: она увеличивает вес более частотных слов и снижает вес менее частотных.



а) б)

Рис. 6 – Дискурсивные маркеры со значением причины (а) и слова категории «Я» (б) в текстах мужчин и женщин

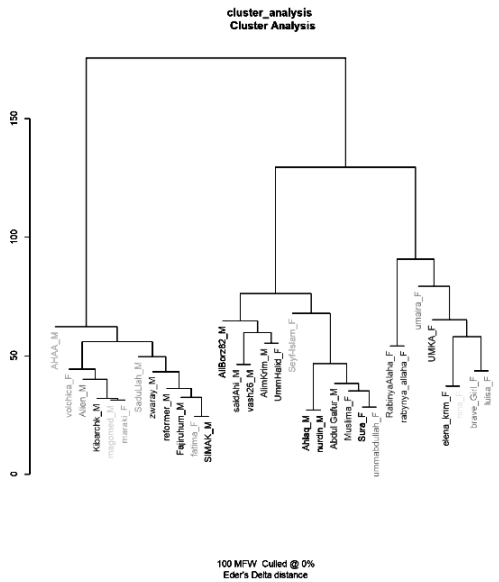


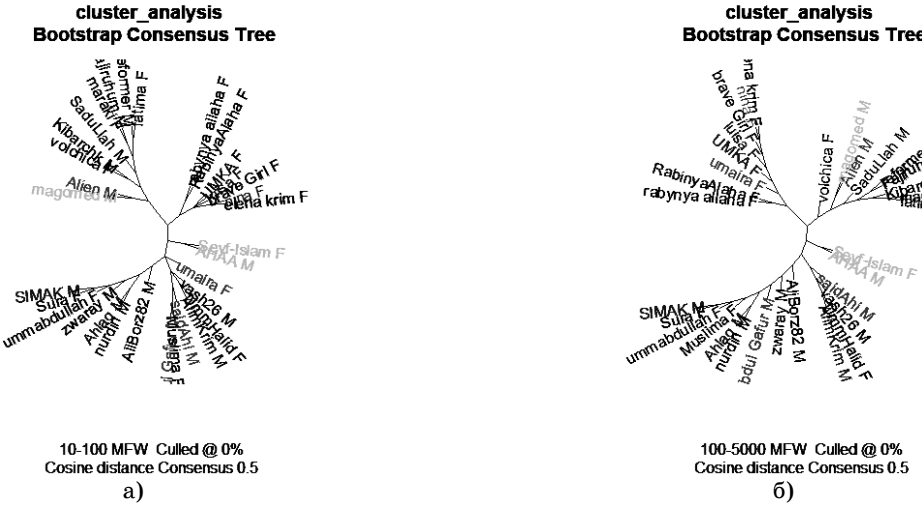
Рис. 7 – Результаты кластерного анализа с использованием дельты Эдера

Проведенный нами анализ выявил тенденцию группировки текстов по половой принадлежности авторов, однако наряду с кластерами женщин и мужчин выделяются и смешанные кластеры. Известно, однако, что кластерный анализ отличается нестабильностью результатов в зависимости от числа признаков, поэтому нами была применена процедура построения консенсусных деревьев, которая рекомендуется в качестве одного из методов решения указанной проблемы [12].

Принцип такой процедуры состоит в построении ряда виртуальных кластеров с разным числом параметров. На консенсусных деревьях отображаются только те связи между текстами, которые встречаются не менее чем в заданном числе кластеров, определяемом силой связи (диапазон значений данного параметра – от 0,4 до 1, то есть отображаются связи, встречающиеся в 40-100% кластеров) [12].

На рис. 7 представлено консенсусное дерево, отображающее 50 % связей, выделенных в серии кластерного анализа для 10-100 самых частотных слов корпуса с шагом 10, то есть преимущественно для строевых слов. Как и в случае с кластерным анализом для 100 самых частотных слов (рис. 6), на консенсусном дереве отчетливо видна группа авторов-женщин (верхняя правая ветвь), авторов-мужчин (левая верхняя ветвь; в нее также входит один автор-женщина, Fatima, как и в случае с кластерным анализом), а также смешанные группы. Таким образом, уже на уровне строевых слов прослеживаются различия между текстами женщин и мужчин, однако данная тенденция характерна не для всех авторов.

Мы повторили эксперимент для более низких уровней частотного списка (состоящих преимущественно из полнозначных слов) с использованием различных мер расстояния, однако, как показывает рис. 8, структура групп практически не изменилась: по-прежнему отчетливо выделяются «женские» (левая верхняя ветвь), преимущественно «мужские» (правая верхняя ветвь) и смешанные группы, причем состав групп авторов, выделенных при помощи разных типов признаков (из верхнего частотного списка, то есть преимущественно строевых, и из нижнего, то есть преимущественно отражающих тематический уровень текста), является схожим.



а) б)

Рис. 8 – Консенсусное дерево для 10-100 самых частотных слов корпуса (а) и 100-5000 самых частотных слов корпуса

Поскольку в нашем корпусе содержится разное число слов от каждого автора, следующим этапом стало разделение текстов автора методом случайного сэмплирования, реализованным в Stylo. Из каждого текста случайным образом отбирались 3 сэмпла длиной 1000 слов, после чего было построено консенсусное дерево для тех же параметров, что и в предыдущем эксперименте, с разными мерами расстояния. На рис. 9 представлен пример консенсусного дерева, построенного с использованием косинусового расстояния, хотя при использовании других мер были получены аналогичные результаты. Мы видим, что при наличии нескольких образцов текстов от одного автора «идиолектный» сигнал оказывается сильнее гендерного: в большинстве случае образцы текстов одного автора образуют одну группу (т.е. оказываются расположенными на одной ветви). Вывод о большей силе авторского сигнала в сравнении с гендерным, возрастным и т.д. сделан также в работе, выполненной на материале англоязычной художественной литературы [13].

Итак, нами было обнаружено, что тексты авторов объединяются по гендерному признаку при использовании в качестве параметров частот слов из разных диапазонов частотного списка, однако указанное наблюдение справедливо не для всех авторов. Заметим, что при проведении описанного выше анализа мы не пользовались методами отбора признаков. На следующем этапе мы поставили цель провести кластерный анализ, используя признаки, различающие авторов – мужчин и женщин. Данные признаки были выделены при помощи функции oppose пакета Stylo. Все тексты были разделены на отрезки по 1000 токенов, в качестве признаков отбирались токены с минимальной встречаемостью в корпусе, равной 10; в качестве меры использовалась Зета Крейга [12] с минимальным значением критерия 0,1. Преимущество данного критерия состоит в том, что он помогает выявить именно

типичные для всей группы текстов особенности, а не присущие отдельным авторам. Результаты, полученные при помощи данной функции, представлены на рис. 10а.

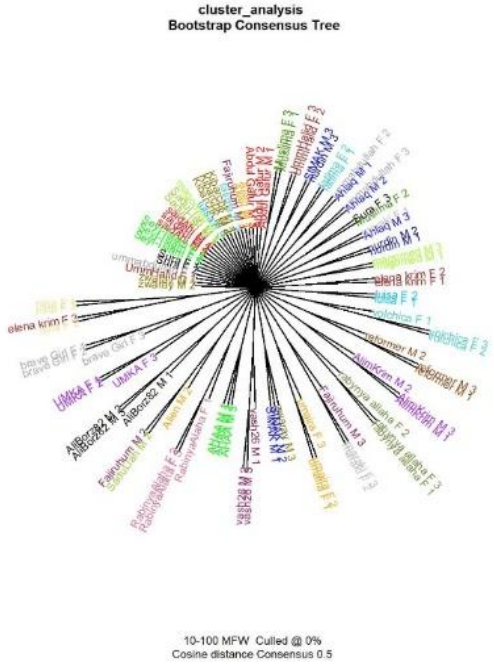
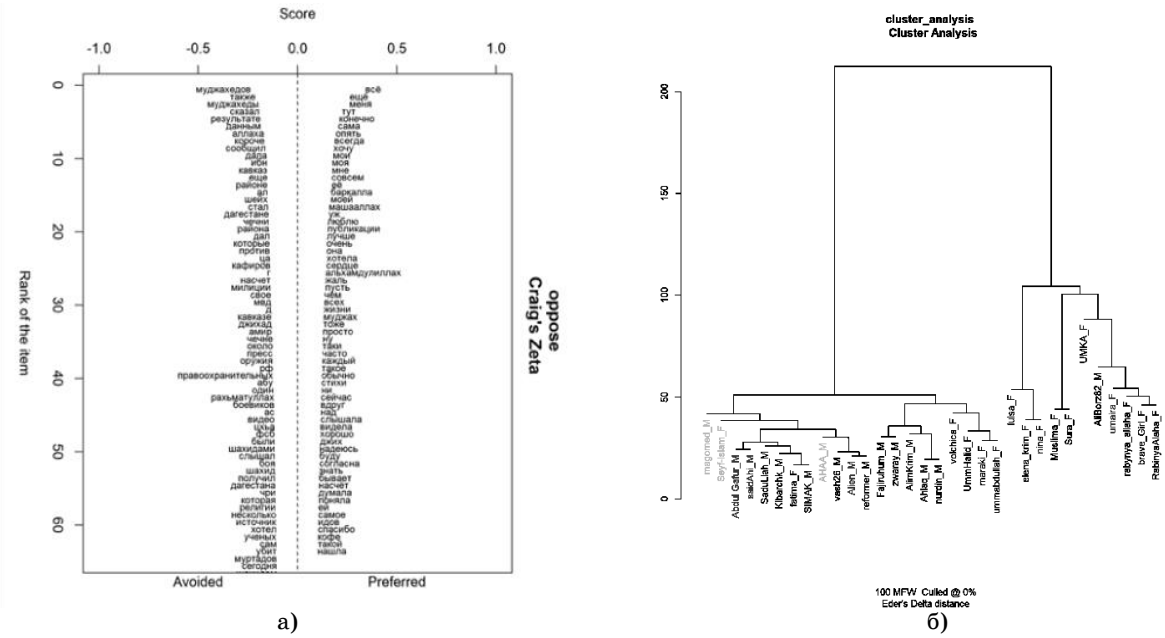


Рис. 9 – Консенсусное дерево на сэмплированном корпусе (число сэмплов = 3, длина = 1000 токенов)



а) б)

Рис. 10 – Результаты работы функции oppose (а) и дендрограмма (б), построенная на их основе

Правый столбец показывает слова, «предпочитаемые», а в левом – «избегаемые» авторами-женщинами в сравнении с авторами-мужчинами. Ряд наблюдений согласуется с полученными ранее данными о преимущественном использовании женщинами местоимений я, интенсификатора очень, отдельных слов группы «чувства» (сердце, люблю), литература (стихи, публикации), мужчинами – слов тематической группы «война» (муджахеды, амир, оружие, шахид и т.д.), «правоохранительные органы» (милиция, правоохранительных). Нами был проведен кластерный анализ на основе полученных данных, т.е. в качестве признаков были взяты частоты слов, выделенных с использованием критерия Зета. Результаты кластерного анализа представлены на рис. 9б. На рис. отчетливо виден преимущественно «женский» кластер (правый), а также большой смешанный кластер, который делится на ряд более мелких кластеров, образованных преимущественно по гендерному признаку. Таким образом, результаты кластерного анализа подтверждают валидность результатов, полученных с использованием критерия Зета.

Для дальнейшего анализа силы «гендерного» сигнала мы провели эксперименты, направленные на классификацию текстов.

Выборка была разделена на тренировочную (24 автора, равное число мужчин и женщин) и тестовую (8 авторов, равное число мужчин и женщин). В качестве признаков были использованы частоты 100–1000 самых частотных слов корпуса (нормированные с использованием z-scores; при нормировании использовались только данные тренировочной выборки). В качестве метода классификации использовался метод опорных векторов (linear SVM). Точность классификатора (выраженная как процент верно классифицированных текстов) представлена в табл. 3.

Таблица 3 – Классификатор

Число признаков	100	200	300	400	500	600	700	800	900	1000
Точность классификатора	62.5	62.5	62.5	75	75	75	75	75	75	75

Таким образом, для текстов преимущественно большого объема увеличение числа признаков свыше 400 не дает дальнейшего увеличения точности модели.

Поскольку наш корпус не сбалансирован в отношении объема текстов от автора, для экспериментов по классификации мы выбрали схему случайного сэмплирования, описанную выше. От каждого автора было взято равное число текстов (10) длиной 1000 слов. Нами была использована схема 10-фолдовой стратифицированной кросс-валидации. Таким образом, 240 текстов общим объемом 240000 токенов использовалось для обучения и валидации, 80 текстов (80 000 токенов) – для тестирования. В качестве признаков использовались 100 – 500 самых частотных слов корпуса. В качестве метода классификации использовался метод опорных векторов (табл. 4).

Таблица 4 – Метод опорных векторов

Число признаков	100	200	300	400	500
Точность классификатора	52.5	76.25	70	71.25	66.25

Примечательно, что с использованием 200 признаков достигается наиболее значительная точность моделей для текстов сравнительно небольшой (1000 токенов) длины.

При проверке результатов классификации нами было выявлено, что разные фрагменты текстов одних и тех же авторов классифицировались неверно при использовании разного числа признаков. На наш взгляд, указанное обстоятельство может свидетельствовать об устойчивости авторского сигнала, а также о гендерной «нетипичности» языковых выборов этих авторов.

На следующем этапе на датасете без сэмплирования мы провели эксперимент с использованием поэлементной кросс-валидации (leave-one-out cross-validation), при которой каждый текст корпуса выступает как тестовая выборка, остальные тексты – как тренировочная, с 1000 самыми частотными словами. Были применены различные алгоритмы классификации, реализованные в пакете Stylo. Наилучшая средняя точность классификации составила 0.7188 (Nearest shrunken centroid). Также нами были исследованы признаков, наиболее значимые при классификации. Анализ признаков показал, что среди наиболее значимых признаков – частоты местоимений я (у женщин показатель выше), мои, она, ничего; женщины также чаще используют частицы же, лишь, ни; лексемы сердце, люблю, наречия всегда, лучше, конечно, очень, союз чтоб. Мужчины чаще используют местоимения их, они, частицу не, глагол будут, предлоги (кроме у, к, со), словоформы оружия, кавказ, религия, мусульмане, русни, джихад, русские.

В таблице 5 приведен пример типичного набора параметров в порядке убывания значимости (первые 20).

Таблица 5 – Набора параметров в порядке убывания значимости

Рейтинг признака	Признак	F-score	M-score	Рейтинг признака	Признак	F-score	M-score
1	меня	0.399	-0.3741	11	из	-0.2026	0.1899
2	я	0.3542	-0.3321	12	по	-0.1855	-0.1855
3	мне	0.3036	-0.2846	13	она	0.1833	-0.1718
4	ни	0.2676	-0.2509	14	в	-0.1781	0.1637
5	к	0.2476	-0.2322	15	их	-0.1746	0.1637
6	на	-0.2388	0.2238	16	все	0.1724	-0.1616
7	хочу	0.2283	-0.2141	17	моей	0.1695	-0.1589
8	этой	-0.2194	0.2057	18	сердце	0.1691	-0.1586
9	у	0.2111	-0.1979	19	ничего	0.1631	-0.1529
10	всегда	0.2039	-0.1911	20	лишь	0.158	-0.1482

Таким образом, проведенные эксперименты показывают наличие различий в лексическом выборе женщин и мужчин – участников экстремистского форума, отобранных при помощи единых формальных критериев и выражающих радикальные взгляды. Следует отметить, что различия обнаруживаются как на уровне строевых слов, так и уровне полнозначных слов, однако степень выраженности ген-

дерно обусловленных различий в лексическом выборе неодинакова для разных авторов: тексты некоторых авторов сложно верно атрибутировать в гендерном аспекте. Несмотря на то, что мужчины, как правило, более сосредоточены на темах оружия, обсуждения России и русских, правоохранительных органов, а женщины – на обсуждении литературы, выражении чувств, данное наблюдение отражает общую тенденцию, но не всегда применимо к отдельным авторам.

Увеличение числа признаков не всегда приводит к росту точности классификатора. Кроме того, так как не все авторы, как было указано выше, являются гендерно типичными в своем речевом поведении, особую значимость приобретает анализ строевых слов. Такие слова наименее подвержены имитации, что особенно важно для установления гендерной принадлежности автора экстремистского текста.

На наш взгляд, указанное исследование доказывает необходимость пересмотра взглядов некоторых авторов, согласно которым автором экстремистского текста является мужчина. В условиях стремительного развития интернет-коммуникации, появления новых форм текстов, в том числе содержащих противоправный контент, проблема диагностической экспертизы авторов таких текстов, в том числе определение их половой принадлежности, приобретает особую актуальность.

Выводы. Проведенное исследование позволяет сделать вывод о перспективности использования стилеметрических методов анализа языкового материала для получения новых данных об особенностях речевого портрета участников экстремистского форума в гендерном аспекте.

Безусловно, выполненное исследование имеет ряд ограничений, связанных преимущественно с небольшим числом авторов, а также отсутствием достоверной информации о половой принадлежности авторов. Последнее ограничение, однако, в целом характерно для работ в области профилирования авторов интернет-текстов.

Очевидно, в дальнейшем необходимо увеличение числа авторов, расширение корпусного материала, исследование новых интернет-жанров для получения более целостного представления об особенностях речевого портрета участников интернет-коммуникации, поддерживающих и/или пропагандирующих радикальные взгляды.

СПИСОК ЛИТЕРАТУРЫ:

1. О проблеме выявления экстремистской направленности в текстах [Текст] / М. И. Ананьева [и др.] // Вестник НГУ. Серия. Информационные технологии. – 2016. – №4. – С. 5-13.

2. Араева, Л. А. Языковая личность экстремиста (о специфике автороведческой экспертизы по криминальным проявлениям экстремизма) [Текст] / Л. А. Араева, М. А. Осадчий // Юрислингвистика. – 2008. – № 9. – С. 182-194.

3. Громова, Н. С. Конституирующие признаки экстремистского дискурса [Текст] / Н. С. Громова // Политическая лингвистика. – 2017. – № 5. – С. 241-245.

4. Литвинова, Т.А. Идентификация и диагностирование автора письменного текста [Текст] / Т.А. Литвинова, О.А. Литвинова. – Воронеж : Изд-во ВГПУ, 2015. – 332 с.

5. Лингвистическая модель диагностики суицидального поведения [Текст] / Т.А. Литвинова [и др.] // Научные доклады высшей школы. – 2017. – № 5. – С. 49-54.

6. Литвинова, Т.А. Лингвистические методы выявления в Сети экстремистского контента и лиц, склонных к экстремизму [Текст] / Т.А. Литвинова, О.В. Загоровская // Современное право. – 2016. – № 3. – С. 107-113.

7. Литвинова, Т.А. Построение речевого портрета участника экстремистского форума с использованием методов корпусной лингвистики [Текст] / Т.А. Литвинова, О.А. Литвинова // Известия Воронежского государственного педагогического университета. – 2019. – № 1. – С. 217-222.

8. Стёпин, Д. С. Особенности осуществления террористической агитации и пропаганды с использованием интернет-ресурсов (на примере форума «Кавказ-Чат») [Текст] / Д. С. Стёпин // Проблемы теории и практики борьбы с экстремизмом и терроризмом : материалы научно-практической конференции. – М. : Российская криминологическая ассоциация; Ставрополь : Изд-во СКФУ, 2015. – С. 25-32.

9. Анализ корпусов текстов террористической и антиправовой направленности [Текст] / А. М. Чеповский [и др.] // Вопросы кибербезопасности. – 2019. – № 4(32). – С. 54-60.

10. Benoit, K. Quanteda: An R package for the quantitative analysis of textual data [Text] / K. Benoit [et al.] // Journal of Open Source Software. – 2018. – № 3(30). – P. 774.

11. De Smedt, T. Multilingual Cross-domain Perspectives on Online Hate Speech. CLiPS Technical Report [Electronic resource] / T De Smedt [et al]. – 2018. – Access mode:https://arxiv.org/abs/1809.03944.

12. Eder, M. Stylometry with R: a package for computational text analysis [Text] / M. Eder, J. Rybicki, M. Kestemont // R Journal. – 2016. – № 8(1). – P. 107-121.

13. Jockers, M. L. Macroanalysis: Digital Methods and Literary History [Text] / M. L. Jockers. – Urbana, Chicago, Springfield : University of Illinois Press, 2013. – 192 pp.

14. Kneip, K. Female Jihad – Women in the ISIS [Text] / K. Kneip // Politikon: IAPSS Political Science Journal. – 2016. – Vol. 29. – P. 88-106.

15. Pennebaker, J. W. Using computer analyses to identify language style and aggressive intent: The secret life of function words [Text] / J. W. Pennebaker // Dynamics of Asymmetric Conflict. – 2011. – № 4(2). – P. 92–102.