УДК 81'322

А.А. Степаненко

ГЕНДЕРНАЯ АТРИБУЦИЯ ТЕКСТОВ КОМПЬЮТЕРНОЙ КОММУНИКАЦИИ: СТАТИСТИЧЕСКИЙ АНАЛИЗ ИСПОЛЬЗОВАНИЯ МЕСТОИМЕНИЙ

Исследование (№ 8.1.37.2015) выполнено при поддержке программы «Научный фонд им. Д.И. Менделеева Томского государственного университета» в 2015–2016 гг.

Представлены результаты применения методов количественного контент-анализа текстов интернет-коммуникации с целью установления гендерных маркеров методами компьютерной лингвистики. Приводятся результаты статистического анализа различий использования местоимений мужчинами и женщинами в компьютерной коммуникации, осуществляется поиск существующих закономерностей их использования в тексте при помощи кластерного анализа. Доказано наличие статистически значимых различий в использовании местоимений Я-группы в текстах мужчин и женщин.

Ключевые слова: компьютерная коммуникация; статистика; атрибуция текста; гендер.

Атрибуция художественного текста – одно из наиболее интенсивно развивающихся направлений в лингвистике, которое сформировалось 40 лет назад и использует методы различных областей науки: лингвистики, логики и математики. Одним из результативных методов атрибуции текста является формально-количественный метод. К числу недавних исследований, в которых применялся данный метод, относится анализ известных классических произведений: «Гарри Поттер» Дж. Роуллинг, «Убить пересмешника» Харпер Ли. Анализ текстов серии произведений Дж. Роуллинг «Гарри Поттер» с использованием контент-анализа позволил автору на основе фиксации количественных показателей употребления языковых единиц подтвердить единое авторство всех произведений [1. С. 25]. Е. Гамерман, изучив на основе контент-анализа частот местоимений и предлогов роман Ли Харпер «Убить пересмешника», приходит к заключению, что произведение, возможно, изменялось и корректировалось редактором. В ходе анализа текста исследователь использовал специализированный пакет stylo в языке программирования R, который позволил визуализировать данные в виде деривационного анализа исследуемых единиц [2. С. 185–186].

Обычно стилеметрические исследования включают несколько этапов: (I) сложная многоуровневая первичная обработка текста, (II) выявление признаков, релевантных для определения авторского инварианта, (III) статистический анализ использования релевантных языковых единиц в тексте, (IV) интерпретация и представление результатов.

Современная лингвистика владеет широким кругом методов стилеметрических исследований текстов, при этом главной проблемой выбора методов анализа и проведения стилеметрического исследования является определение начальных критериев и признаков, по которым будет проводиться количественный анализ. На данный момент уже выявлен ряд лексических и синтаксических критериев для анализа идиостиля, появляются новые и модифицируются старые критерии атрибуции текста (см., например, об этом [3. С. 31; 4. С. 42], а также [5. С. 279; 6. С. 20; 7. С. 86; 8. С. 246]). Как показывает анализ практики применения различных методов атрибуции текста, при их реализации исследователи в качестве маркеров опираются на определение частотности использования единиц синтаксического, лексико-фразеологического, стилистического уровней, значимыми признаются также признаки пунктуационного, орфографического аспектов текста. Использование в качестве маркеров единиц синтаксического и лексического уровней наиболее распространено в практике стилеметрического анализа и дает более точные результаты. Однако следует подчеркнуть, что исследователи, как правило, работают сразу с группой признаков для доказательства или опровержения какой-либо гипотезы относительно авторства текста.

Вторая проблема, с которой сталкиваются исследователи, это выбор конкретных методик. В зависимости от поставленной гипотезы исследователи применяют несколько методов в атрибуции текста, конкретных статистических методик, что позволяет добиться более точного результата. Так, например, при исследовании частот лексических единиц следует учитывать их распределение, которое может быть параметрическим и непараметрическим. На рис. 1 приведены методы проверки статистических гипотез, применяемых в современных стилеметрических исследованиях.

В качестве примера реализации данных методов можно привести программу «СМАЛТ», которая основывается на таких статистических методах, как критерии Колмогорова–Смирнова, Стьюдента и кластерный анализ, используемых при анализе особенностей синтаксических структур текста с использованием деревьев зависимостей и типов связей, деревьев зависимостей и мер сложности [9. С. 155–159].

Более точные результаты дает комбинация статистических методов и метода нейронных сетей. Примером таких исследований являются работы, выполненные с использованием программы В.В. Поддубного и А.А. Поликарпова «Лингвоанализатор», интегрирующей применение в анализе текста марковские цепи, нейронные сети прямого распространения, деревья решений, меры расстояния, например [10. С. 55–56].

В исследованиях, выполненных под руководством Л.В. Милова, атрибуция текстов проводится при помощи построения графов «сильных связей» по матрице частот парной встречаемости грамматических классов слов и осуществляется при помощи специальной компьютерной программы [11. С. 356].

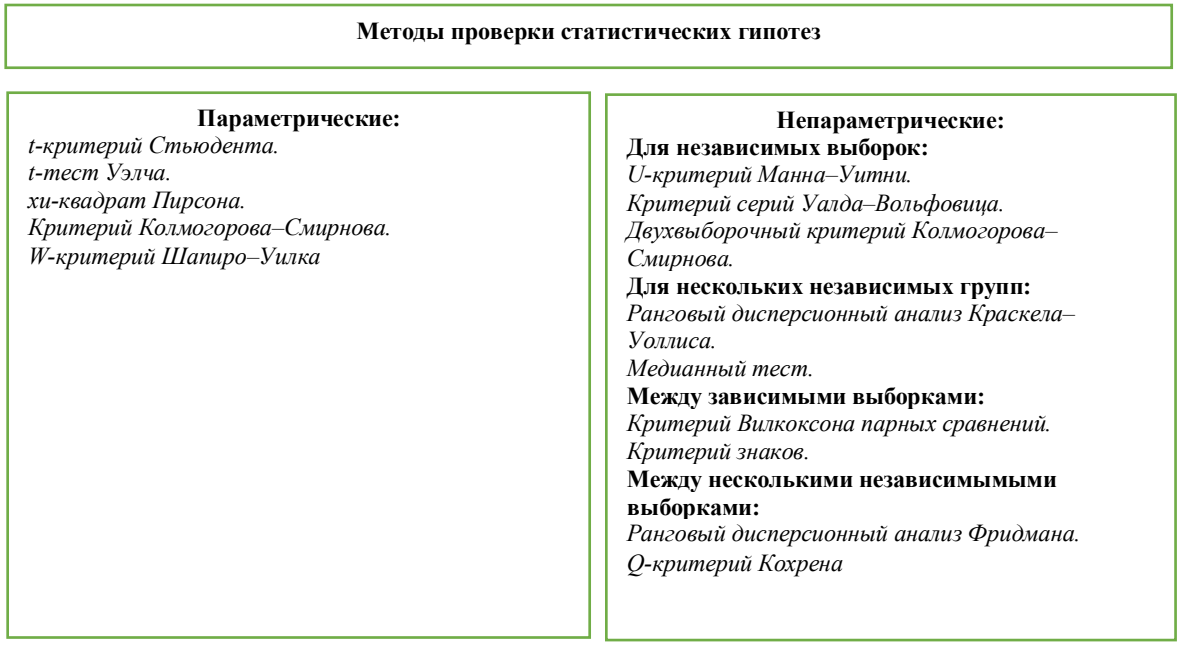

Рис. 1. Методы проверки статистических гипотез

Третья проблема, с которой сталкиваются ученые, – ограничения при отборе материала исследования: это требование к характеру текста (необходимое качество – однородность текстов) и к их объему (исключаются тексты небольшого объема). Стилеметрия как направление научных исследований формировалась и апробировала свои методы на материале художественных произведений, что во многом было мотивировано доступностью объемных текстов однотипной функциональной структуры. Это, в свою очередь, обусловлено тем, что процесс установления авторства формально-количественными методами при исследовании малого объема текста может показывать некорректный результат или вызвать ошибку. Особенно это касается подсчета биграмм символов. Согласно В.П. Фоменко, минимальный объем исследуемых текстов должен составлять не менее 8 000 символов [12. С. 769]. Д.В. Хмелев, использовавший энтропийный метод классификации (с помощью сжатия), основанный на цепях Маркова первого порядка, приходит к заключению, что данный метод показывает хорошие результаты на файлах большого объема и плохие, по сравнению с другими методами, в текстах длиной в 2 000–5 000 символов [13]. Однако следует отметить и программы, которые успешно работают на малом объеме выборки (500 слов), к числу которых относится «СМАЛТ» [9. С. 155–160].

В настоящее время методы, выработанные при анализе художественных произведений, переносятся на анализ текстов других функциональных типов. К числу приоритетных задач, решаемых в этом направлении, следует отнести установление авторства текстов компьютерной коммуникации, что мотивировано как мощным развитием в настоящее время этого средства коммуникации, опосредствующей все виды социальных практик современного человека, так и большими возможностями фальсификации авторства таких текстов по сравнению с другими типами коммуникации [14. С. 727–728].

В работах, посвященных атрибуции компьютерно созданных текстов, исследуются способы обнаружения как индивидуального автора (Т.А. Литвинова, О.С. Поршнева, Х. Хьетсо) [15. С. 196–197; 16. С. 38–39; 17. С. 182–188], так и модели типовых авторов, выделенных по каким-либо социально значимым параметрам (Т.Н. Дроздова, А.С. Романов) [18. С. 400–401; 19. С. 26].

В ряду последних следует отметить и работы, в которых моделируются типовые авторы, противопоставленные по гендерному признаку. Большинство исследований компьютерной коммуникации, посвященных гендерной принадлежности автора текста, опираются на методологию атрибуции художественного текста, ограничивая параметры сравнения единицами синтаксического уровня [20. С. 27–28; 21. С. 33–35; 22. С. 23–26].

Подобную проблему решали лингвисты в сравнении выборок художественных и научных текстов. К таким исследованиям можно отнести работу Ш. Аргомона (Sh. Argamon) и М. Коппела (M. Koppel). Авторы, обратившись к материалам Британского национального корпуса, включающего в себя большой выбор жанров и используя метод нейронных сетей, установили различия между мужской и женской письменной речью. Ученые выявили классы лексических и синтаксических особенностей, отличающиеся в коммуникации мужчин и женщин и идентифицирующиеся в научных и художественных текстах. В частности, отмечаются существенные различия в использовании местоимений и некоторых существительных в мужских и женских текстах: женщины используют больше местоимений, а мужчины – больше специфичных существительных [23. С. 715–720]. Однако авторы подчеркивают, что применение тех же методов в исследованиях художественных текстов, направленных на поиск индивидуального стиля автора, дает более точные результаты, повышая вероятность принятия гипотезы до 98%,

нежели при анализе текстов компьютерной коммуникации. Авторы делают вывод, что методы, применяемые в атрибуции компьютерной коммуникации, направленные на установление общих характеристик определенной группы авторов мужского или женского пола, имеют более высокую погрешность, нежели применяемые при определении индивидуальных особенностей стиля автора художественного текста. Данная, ещё не решённая проблема ставит под сомнение корректность «прямого» переноса методов стилеметрии при атрибуции художественного текста на идентификацию гендерных субъектов компьютерной коммуникации и требует изменений в применении формально-количественных методов. При этом авторами по отношению к материалам данного, весьма специфического типа коммуникации также решаются и другие отмеченные выше проблемы применения стилеметрических методов: проблема выбора единиц анализа, проблема ограничений объема текста и его однородности [24. С. 25–26].

М. Брокардо (M. Brocardo) предложил решение проблемы ограниченных исходных данных при установлении гендерной принадлежности автора текста в компьютерной коммуникации (общение в «Твиттере») на основе подсчета n-грамм (2-грамм, 3-грамм, 4-грамм, 5-грамм) символов и с использованием метода нейронных сетей [25]. Исследование отличается тем, что анализируемые сообщения не превышают 500 лексических единиц. Задачей исследователя является поиск наиболее точного определения автора, на основе разработанного алгоритма вероятных ошибок, возникающих в условиях малой выборки при поиске биграмм. Определяются три группы вероятных ошибок, позволяющих оптимизировать процесс установления гендера автора в компьютерной коммуникации:

1. False Acceptance Rate (FAR) – мера, в которой система неверно определяет истинного автора текста.

2. False Rejection Rate (FRR) – мера, где алгоритм не распознает автора сообщения.

3. Equal Error Rate (ERR) – точка, в которой FAR и FRR принимает равное значение и увеличивает точность определения автора анализируемого текста. Исследователь приходит к заключению, что любой частотный анализ n-грамм вызывает появление одной из вышеуказанных ошибок, но наиболее точными, в условиях ограниченной выборки, являются 5-граммы, где FRR = 14,71%, FAR = 13,93%.

В своей работе мы учитываем преимущества и недостатки вышеуказанных исследований и применяем формально-количественные методы поиска авторского инварианта, разработанные и используемые М. Аррой (M. Arroju) [26. С. 23–31] в анализе текстов компьютерной коммуникации.

Нашей основной задачей является выявление *групповых гендерных различий авторского инварианта в компьютерной коммуникации.* В качестве объекта исследования мы использовали текст компьютерной коммуникации из социальной сети «ВКонтакте», который представляет собой неформальную коммуникацию мужчин и женщин в возрасте 18–20 лет.

Данное исследование включало несколько этапов:

1. Сбор текстового материала и выборка реплик диалогов компьютерной коммуникации (далее – объекты) по гендерному принципу (мужские и женские).

2. Определение переменных, по которым будут оцениваться объекты в выборке, т.е. поиск признакового пространства.

3. Определение того, существуют ли статистически значимые различия между двумя независимыми группами по отобранным признакам.

4. Подсчет частоты лексических единиц в качестве значений признаков, составляющих вектор, для проведения неиерархического метода кластерного анализа (k-средних) в объектах исследования.

5. В качестве вывода выносим альтернативные гипотезы:

A. H_0 – выбранные группы не имеют значимых различий по исследуемому признаку.

B. H_1 – выбранные группы значимо различаются по исследуемому признаку.

Если эмпирическое значение равно или превышает теоретическое значение критерия, то отклоняем гипотезу H_0 и принимаем гипотезу H_1. Для расчета критериев применяем специализированные пакеты обработки статистических данных STATISTICA.

На первом этапе нами было собрано 19 диалогов *личных сообщений* (*межличностной коммуникации*) «Мужчина – Женщина». Общее количество информантов – 38 человек в возрасте 18–20 лет. Размер каждого диалога составил 150–200 Кб (Один диалог – около 10 страниц печатного текста). Все тексты были собраны в рамках учебной практики студентов отделения Фундаментальной и прикладной лингвистики Томского государственного университета, выполняемой под руководством автора статьи. Извлечение диалогов из социальных сетей осуществлялось с согласия их авторов, которые, в соответствии с нормами регламента Этического комитета междисциплинарных исследований ТГУ (http://lab.tsu.ru/ cognitivestudies/node/14) и в соответствии с Федеральным законом № 152 РФ «О персональных данных», были проинформированы о целях проводимого исследования и о гарантиях анонимности предоставленных персональных данных, после чего были заполнены «Формы информированного согласия», в структуру которых были включены метаданные участников диалогов: пол, возраст, социальный статус. Для дальнейшей статистической обработки все диалоги были разделены на файлы, содержащие мужские и женские реплики по 49–50 Кб каждая.

В исследуемых текстах нашей задачей является поиск характерных стилистических особенностей, гендерных маркеров стиля речи, принадлежащих мужчинам или женщинам.

На данном этапе были получены также общие количественные данные о мужских и женских репликах диалогов (таблица). Как видно из таблицы, в процессе коммуникации при одинаковых условиях женщины используют больше слов, однако, средняя длина предложения значительно меньше, чем у мужчин.

Следующий этап исследования заключался в поиске обоснованных переменных для проведения статистических исследований.

Данные о мужских (М) и женских (Ж) репликах диалогов

Показатель	Ж	М
Всего букв	214 829	165 667
Всего слов	47 966	37 988
Всего предложений	3 161	1 771
Средняя длина слова	4,479	4,361
Средняя длина предложения	15,174	21,450

В данной работе мы, поставив задачу выявления различий в принципах строения мужского и женского текста, ограничиваемся анализом использования местоимений. Основой выбора исследования местоимений послужили работы отечественных и зарубежных психологов, посвященных «Я-концепции» и самосознания в контексте общего развития личности. С точки зрения Л.С. Рубинштейна, «самосознание – осознание себя субъектом деятельности, сложное, интегративное, прижизненно формирующееся свойство психической деятельности личности, осознание собственных действий, результатов, поступков, мыслей, мотивов, ценностей, оценка себя и своего места в жизни» [27. С. 244].

Данная сложная структура психической деятельности объективируется в речевых практиках, в которых значимую роль играет позиционирование говорящим себя по отношению к другим участникам коммуникации. В лингвистических исследованиях местоимений учеными было доказано, что этот класс лексических единиц служит одним из значимых средств маркирования позиции говорящего по отношению к другим коммуникантам (Дж. Пеннебекер [28. С. 563–565], Е.М. Вольф [29. С. 112] и др.). В работах же по гендерной лингвистике обыло отмечено различие в использовании местоимений мужчинами и женщинами. Так, Б. Верховен (Ben Verhoeven) доказывает на материале мультилингвальных электронных корпусов текстов, что женщины используют местоимение «Я» чаще, чем мужчины [30. С. 1632–1633]. Подобное исследование проводилось А.Н. Барановым в статистическом анализе использования местоимений в художественных текстах [31. С. 235].

Вслед за вышеуказанными исследователями мы выдвигаем гипотезу, что использование местоимений в текстах коммуникации социальных сетей может быть маркером гендерных различий коммуникантов и выражает их «Я-позицию» в процессе коммуникации относительно участников диалога.

При анализе местоимений в текстах мы исходим из того, что местоимения имеют, во-первых, собственное внеконтекстное постоянное значение и, во-вторых, контекстуальное (ситуативное) значение, определяемое дейктической (указательной) функцией. Учет типов ведущей референции местоимений, используемых говорящим в том или ином дискурсе, может, по нашему мнению, свидетельствовать о важных аспектах коммуникативных установок говорящих. Для нас в данном аспекте важны противопоставление направленности субъектов коммуникации на себя или ориентация на партнера по коммуникации. Личные местоимения выступают одним из ярких показателей эгоцентричной или партнерской направленности наряду с другими словообразовательными, лексическими и синтаксическими средствами. Наиболее ярким маркером выражения эгоцентризма в речи является соотношение групп местоимений, производных супплетивных форм «я» (первого лица единственного числа) – мой, мое, моя, мне и т.д. (далее условно их будем именовать «группой местоимений первого лица единственного числа от “Я”»), а также представляет интерес позиция говорящего, выраженная в системе других местоимений.

Для проверки сформулированной гипотезы мы провели анализ частоты употребления местоимений в текстах нашей выборки.

Результат подсчета употреблений всех групп местоимений в совокупности начальных и косвенных падежных форм представлен в следующей диаграмме, отражающей среднее значение (mean) абсолютных чисел местоимений, использованных мужчинами (м) и женщинами (ж) в проанализированных текстах (рис. 2).

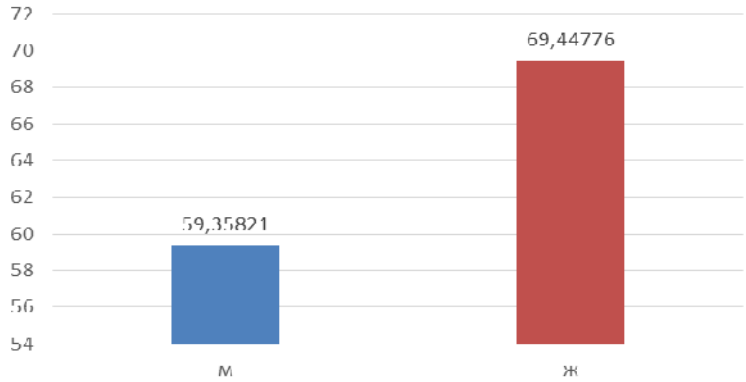

Рис. 2. Среднее значение (mean) абсолютных чисел использования местоимений в проанализированных текстах

Диаграмма абсолютных чисел доказывает, что женщины чаще используют местоимения в процессе коммуникации (рис. 2). Однако анализ показателей использования местоимений относительно общего количества слов проанализированных текстов показал противоположенный результат: мужчины чаще используют местоимения в тексте (рис. 3).

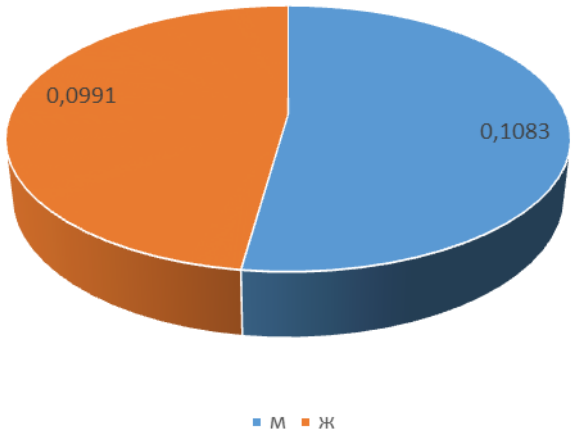

Рис. 3. Соотношение количества использованных местоимений и общего количества слов текста

Анализ относительных переменных также указывает на статистически значимые различия в использовании местоимений мужчинами и женщинами относительно общего числа слов в тексте. Учитывая экспоненциальную функцию распределения местоимений (Критерий $x^2 = 3{,}477$; $p = 0{,}324$), для проверки гипотезы был применен тест Уалда–Вольфовица. Изучив и проанализировав результат теста, мы приняли альтернативную гипотезу (H_1), где $Z = 1{,}97$; $p = 0{,}048$. Дальнейшим шагом анализа стала верификация различий в частоте использования местоимений мужчинами и женщинами относительно всех слов текста. Дисперсионный анализ ANOVA не выявил статистически значимых различий, что требует более детального изучения (рис. 4). Однако, как можно видеть на графике, совокупное количество использований местоимений всех групп хоть и незначительно, но преобладает в мужской коммуникации.

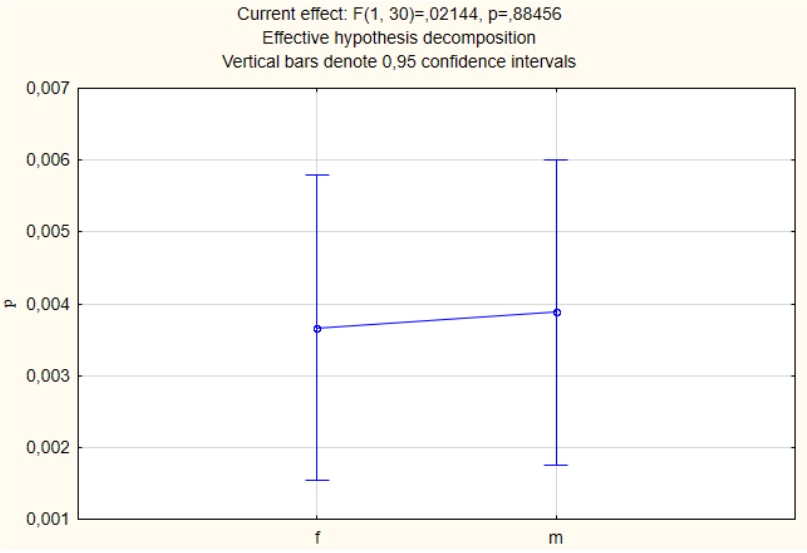

Рис. 4. Частота использования местоимений мужчинами и женщинами относительно общего числа слов в тексте

Учитывая значительные прагматические различия использования в коммуникации местоимений разных лексико-грамматических разрядов, мы выделили группы местоимений, противопоставленные по типу маркирования позиций коммуникантов в речи, и проанализировали частотность их употреблений в анализируемых текстах:

1. «Я-группа»: все словоформы местоимения 1-го лица ед. ч., а также словоформы притяжательного местоимения *мой*.

2. «Ты-группа»: все словоформы местоимения 2-го лица ед. ч., а также словоформы притяжательного местоимения *твой*.

3. «Мы-группа» – все словоформы местоимения 1-го лица мн. ч., а также словоформы притяжательного местоимения *наш*.

4. «Сам-группа» – все формы возвратных местоимений.

Для проверки значимости различия между средними в разных группах с помощью сравнения дисперсий, мы применили анализ ANOVA. Результаты анализа представлены на рис. 5. Знак * обозначает статистически значимое различие, M-F m – количество используемых местоимений соответствующей группы в текстах мужчин, M-F f – количество используемых местоимений в текстах женщин.

Как видно из рис. 5, обнаруживаются различия в распределении личных местоимений только «Я-группы» по отношению ко всем другим группам личных местоимений в текстах мужчин и женщин. Однако данные о частотности использования личного местоимения единственного числа свидетельствуют о том,

что в общении женщины возрастной группы 18–20 лет чаще используют группу местоимений «Я», в то время как использование местоимений других функционально-семантических групп не выявляет статистически значимых различий. Данные показатели говорят об эгоцентричной коммуникации женщин, а также эксклюзивности, т.е. направленности в процессе коммуникации на себя.

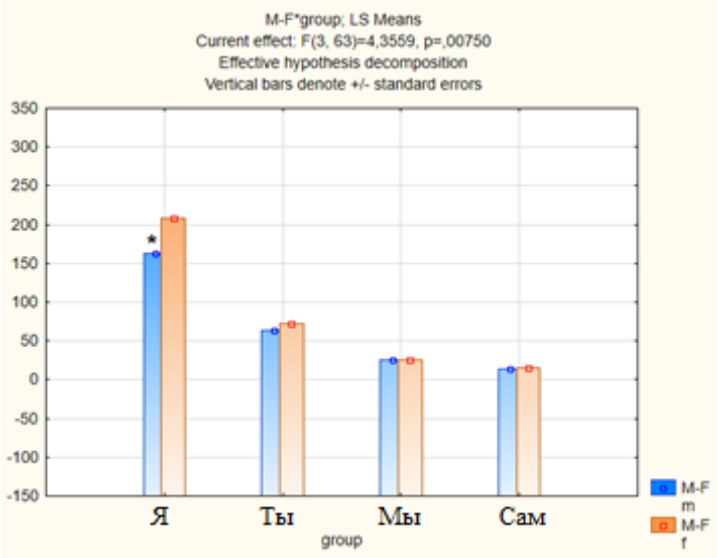

Рис. 5. Различия в использовании местоимений четырех функционально-семантических относительно общего числа слов в тексте

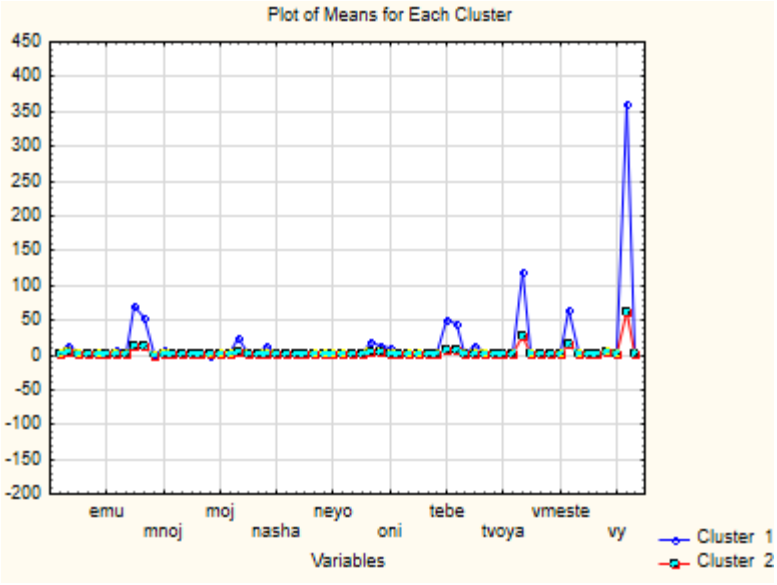

Рис. 6. Кластерный анализ местоимений

Достоверность результата была подтверждена при помощи многофакторного дисперсионного анализа ANOVA в коммуникативных группах местоимений. Анализ подтверждает статистически значимые различия в абсолютных числах в супплетивной группе местоимений «Я». P-значение удовлетворяет условию $p < 0{,}05$, где $F = 4{,}36$; 1; $p = 0{,}0075$, что позволяет принять гипотезу H_1. Для подтверждения гипотезы мы провели дополнительный статистический анализ полученных данных. Во-первых, мы установили, что распределение является непараметрическим и, во-вторых, выбрали наиболее подходящий критерий проверки гипотезы – U-критерий Манна–Уитни, который используется для оценки различий между двумя независимыми выборками (мужчины и женщины) по уровню какого-либо признака (местоимения), измеренного количественно. Вычисленный критерий составил $p = 0{,}0169$. Так как вычисленное значение критерия меньше установленного 0,05, нулевая гипотеза (H_0) отвергается на выбранном уровне значимости и различия между выборками признаются статистически значимыми. Таким образом, вывод о существовании различий, сделанный с помощью непараметрического критерия Манна–Уитни, подтверждается с помощью данного непараметрического метода, а значит, нами установлено, что имеются существенные различия использования мужчинами и женщинами местоимений в компьютерной коммуникации.

Валидность эксперимента подтверждает также однофакторный дисперсионный анализ ANOVA, сравнивающий выборку употребления средних показателей местоимений ($p = 0{,}0252$), что также позволило отклонить нулевую гипотезу о равенстве дисперсий в изучаемых группах. Р-значение удовлетворяет условию $p < 0{,}05$, где $F = 7{,}53$; 1; $p = 0{,}025$.

Проведение кластерного анализа для выявления различий в использовании местоимений в исследуемом типе текстов мужчинами и женщинами не дало положительного результата. Исходные данные нашей задачи были представлены также в виде строковых типов переменных по 49–50 Кб отдельно попарно для женской и мужской коммуникации, хранимой в файле Microsoft Word. Данные были импортированы в среду STATISTICA, где подверглись предварительной обработке. Обработка заключалась в удалении всех лексических единиц, кроме местоимений. Выяснилось, что кластеры присваивались в зависимости от стиля и объема диалога между коммуникантами, но эти данные нестабильны и требуют дальнейшего исследования, а использование местоимений не было проинтерпретировано в качестве релевантно значимых признаков (см. рис. 6).

Таким образом, проведение статистического анализа использования местоимений в компьтерно-опосредствованной коммуникации (общение в социальной сети «ВКонтакте») позволяет сделать выводы об отсутствии статистически значимых различий в использовании местоимений в совокупности всех разрядов данной части речи, но выявляется значимое статистическое различие в использовании местоимений Я-группы, что согласуется с ранее сделанными в гендерной лингвистике выводами о более выраженной эгоцентричной направленности женской коммуникации, однако необходимо выявление групп лексики других лексико-грамматических классов, способных выявить данный аспект коммуникативных стратегий.

ЛИТЕРАТУРА

1. Patrick Juola. How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling. Author of the Harry Potter books has a distinct linguistic signature // Scientific American. 2013. P. 24–29.
2. Maciej Eder, Jan Rybicki. Go Set A Watchman while we Kill the Mockingbird in Cold Blood, with Cats and Other People // Digital Humanities. Krakow, 2016. P. 184–186.
3. Мартыненко Г.Я. Основы стилеметрии. Л. : Изд-во Ленинград. ун-та, 1988. 173 с.
4. Резанова З.И., Романов А.С., Мещеряков Р.В. О выборе признаков текста, релевантных в автороведческой экспертной деятельности // Вестник Томского государственного университета. Филология. 2013. № 6 (26). С. 38–52.
5. Аверьянов Л.Я. Контент-анализ. М. : Изд-во РГИУ, 2007. 456 с.
6. Антонова И. Анализ количества деловой информации как инструмент маркетинга // Информационные ресурсы России. 2003. № 5. С. 18–20.
7. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л. : Изд-во Ленинград. ун-та, 1990. 164 с.
8. Гилилов И.М. Игра об Уильяме Шекспире, или Тайна Великого Феникса. М. : Междунар. отношения, 2007. 536 с.
9. Рогов А.А., Гурин Г.Б., Котов А.А., Сидоров Ю.В., Суровцова Т.Г. Программный комплекс СМАЛТ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : тр. X Всерос. науч. конф. «RCDL'2008». Дубна, 2008. 160 с.
10. Поликарпов А.А., Поддубный В.В., Кукушкина О.В., Кубарев А.И., Варламов А.А., Суровцева Е.В., Пирятинская Е.Ф. Комплексная тексто-аналитическая система «СтилеАнализатор-2», основанная на Web-технологиях: разработка, наполнение данными и тестирование на прикладных задачах. М., 2013. 66 с.
11. Милов Л.В. От Нестора до Фонвизина. Новые методы определения авторства. М. : Прогресс, 1994. С. 356.
12. Фоменко В.П. Авторский инвариант русских литературных текстов // Фоменко В.П., Фоменко Т.Г. Новая хронология Греции. Античность в Средневековье. М. : Изд-во Учебно-научного центра довузовского образования Моск. гос. ун-та, 1996. Т. 2. С. 820.
13. Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое ведение. URL: http://compression.graphicon. ru/download/articles/classif/intro.html, свободный (дата обращения: 16.09.2016).
14. Azarbonyad H. Time-Aware Authorship Attribution for Short Text Streams // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. N.Y., 2015. P. 727–730.
15. Литвинова Т.А. Русский письменный текст как носитель информации об индивидуально-личностных характеристиках его автора (на материале корпуса текстов нового типа Personality) // Известия ВГПУ. Сер.: Педагогические науки; Гуманитарные науки. 2015. Т. 266, № 1. С. 196–198.
16. Поршнева О.С. К вопросу об атрибуции текстов записей солдатских разговоров // Информационный бюллетень ассоциации «История и компьютер» / отв. ред. Л.И. Бородкин. М., 2002. № 30. С. 31–44.
17. Хьетсо Г. Кто написал «Тихий Дон»? М. : Книга, 1989. 186 с.
18. Дроздова Т.Н. Диагностические и классификационные задачи в автороведческой экспертизе блогов // Актуальные проблемы российского права. 2010. № 2 (15). С. 394–404.
19. Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста : автореф. дис. ... канд. техн. наук. Томск, 2010. 130 с.
20. Мамаев М.М. Гендерная атрибуция переводных текстов как специфический случай исследования языкового сознания автора // Вестник МГОУ. Сер. Лингвистика. 2015. № 2. С. 25–31.
21. Mukherjee A., Liu B. Improving Gender Classication of Blog Authors // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010. P. 32–38.
22. Yan X., Yan L. Gender Classification of Weblog Authors // Computational Approaches to Analyzing Weblogs. AAAI, 2006. P. 18–26.
23. Shlomo Argamon Gender, Genre, and Writing Style in Formal Written Texts // Shlomo Argamon, Moshe Koppel, Jonathan Fine, Anat Rachel Shimoni Springer, Sex Roles. 2010 Jun. № 62 (11–12). P. 705–720.
24. Резанова З.И., Романов А.С., Мещеряков Р.В. Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) // Вестник Томского государтсвенного универистета. 2013. № 370. С. 24–28.
25. Marcelo Luiz. Brocardo Authorship Verification for Short Messages using Stylometry, 2014. URL: https://www.deepdyve.com/lp/institute-of-electrical-and-electronics-engineers/authorship-verification-for-short-messages-using-stylometry-JM5XWbkHyN (дата обращения: 7.07.2016).
26. Arroju M. Age, Gender and Personality Recognition using Tweets in a Multilingual Setting // 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction, 2015. P. 23–31.
27. Рубинштейн С.Л. Основы общей психологии. М. : Педагогика, 1989. Т. 1. 720 с.

28. Pennebaker J.W., MR Mehl K.G. Niederhoffer Psychological aspects of natural language use: Our words, our selves // Annual review of psychology. 2003. P. 548–571.
29. Вольф Е.М. Грамматика и семантика местоимений. М. : Наука, 1974. 223 с.
30. Verhoeven B.C. TWISTY: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling // Ben Verhoeven, Walter Daelemans and Barbara Plank CLiPS Research Center, University of Antwerp, Belgium University of Groningen, The Netherlands, 2015. P. 1632–1637.
31. Баранов А.Н. Введение в прикладную лингвистику. М. : Эдиториал УРСС, 2001. 347 с.

Статья представлена научной редакцией «Филология» 19 января 2017 г.

GENDER ATTRIBUTION IN SOCIAL NETWORK COMMUNICATION: THE STATISTICAL ANALYSIS OF PRONOUNS FREQUENCY

Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal, 2017, 415, 17–25.
DOI: 10.17223/15617793/415/3
Andrey A. Stepanenko, Tomsk State University (Tomsk, Russian Federation). E-mail: stepanenkone@mail.ru
Keywords: network communication; statistics; authorship attribution; gender.

Authorship attribution in literature is one of the rapidly developing areas in linguistics, which was formed 40 years ago. Today it combines different methods of science: linguistics, logic, mathematics. The combination of methods has allowed using their variety, which could increase accuracy in authorship attribution of the text. However, the main problem of this field is connected with choosing initial criteria and indicators during quantitative text analysis. This article describes the modern criteria for the text analysis of markers based on the frequency determination of the syntactic units, phraseological and stylistic levels. Unlike art texts, network communication is not structured and it has a small size. In this aspect, the problem of network text analysis is more focused on the marker identification of group speakers than on author's individual invariant. Therefore, modern research in text attribution needs method transformation taking into account the above problems. In this article, the author adapts quantitative linguistics methods of art text attribution taking into account differences in using gender markers during computer communication. This research consisted of the following stages: 1) collection of text material and grouping of computer communication dialogues by gender (male and female); 2) choice of variables the objects are assessed by in the sample, i.e. search for the attribute space on the gender basis; 3) analysis of statistically significant differences between the two independent groups in selected attributes. The attributes include pronouns as gender markers which express differences in I-positions in communication. To identify gender differences in the expression of I-positions, the author analyzed informal dialogues from the social network VKontakte. All texts represented informal communication between men and women (18–20 y.o.). The total number of respondents was 38 people. The size of each dialogue was about 150-200 KB (one conversation made up 10 printed pages). To find statistically significant values in the dialogues, they were divided into files containing male and female lines (49–50 KB each). Personal pronouns in the texts were marked and classified into several groups. The hierarchical cluster analysis method (k-means) was used as the main method for the research objects. The results of the statistical analysis showed differences in the distribution of personal pronouns in "I-group". However, the frequency of the use of the personal pronoun in the singular suggests that women aged 18–20 use more "I-group" pronouns, while the use of pronouns other functional-semantic groups did not reveal statistically significant differences. These figures showed that women's communication is more self-centered and exclusive. The further object of the search is a quantitative analysis of emotional markers in communication and building a classifier.

REFERENCES

1. Juola, P. (2013) How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling. Author of the Harry Potter books has a distinct linguistic signature. *Scientific American*. August 20. pp. 24–29.
2. Eder, M. & Rybicki, J. (2016) Go Set A Watchman while we Kill the Mockingbird in Cold Blood, with Cats and Other People. *Digital Humanities*. Krakow. pp. 184–186.
3. Martynenko, G.Ya. (1988) *Osnovy stilemetrii* [Basics of stylemetry]. Leningrad: Leningrad State University.
4. Rezanova, Z.I., Romanov, A.S. & Meshcheryakov, R.V. (2013) Selecting text features relevant for authorship attribution. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 6 (26). pp. 38–52. (In Russian). DOI: 10.17223/19986645/26/4
5. Aver'yanov, L.Ya. (2007) *Kontent-analiz* [Content analysis]. Moscow: RGIU.
6. Antonova, I. (2003) Analiz kolichestva delovoy informatsii kak instrument marketinga [Analysis of the amount of business information as a marketing tool]. *Informatsionnye resursy Rossii*. 5. pp. 18–20.
7. Marusenko, M.A. (1990) *Atributsiya anonimnykh i psevdonimnykh literaturnykh proizvedeniy metodami raspoznavaniya obrazov* [The attribution of anonymous and pseudonymous literary works by methods of image recognition]. Leningrad: Leningrad State University.
8. Gililov, I.M. (2007) *Igra ob Uil'yame Shekspire, ili Tayna Velikogo Feniksa* [The Shakespeare Game, or The Mystery of the Great Phoenix]. Moscow: Mezhdunar. otnosheniya.
9. Rogov, A.A. et al. (2008) [SMALT software package]. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii* [Digital Libraries: Advanced Methods and Technologies, Digital Collections]. Proceedings of the Xth all-Russian conference "RCDL'2008". Dubna.
10. Polikarpov, A.A. et al. (2013) *Kompleksnaya teksto-analiticheskaya sistema "StileAnalizator-2", osnovannaya na Web-tekhnologiyakh: razrabotka, napolnenie dannymi i testirovanie na prikladnykh zadachakh* [Integrated text-analytical system "StileAnalizator-2" based on Web-technologies: the development, filling and testing of data on applied problems]. Moscow.
11. Milov, L.V. (1994) *Ot Nestora do Fonvizina. Novye metody opredeleniya avtorstva* [From Nestor to Fonvizin. New methods for authorship attribution]. Moscow: Progress.
12. Fomenko, V.P. (1996) Avtorskiy invariant russkikh literaturnykh tekstov [Author invariant of Russian literary texts]. In: Fomenko, V.P. & Fomenko, T.G. *Novaya khronologiya Gretsii. Antichnost' v srednevekov'e* [New Chronology of Greece. Antiquity to the Middle Ages]. Vol. 2. Moscow: Teaching and Research Center of pre-university education of Moscow State University.
13. Khmelev, D.V. (2003) *Klassifikatsiya i razmetka tekstov s ispol'zovaniem metodov szhatiya dannykh. Kratkoe vvedenie* [Classification and text layout using data compression techniques. Brief introduction]. [Online] Available from: http://compression.graphicon.ru/download/articles/classif/intro.html. (Accessed: 16 September 2016).
14. Azarbonyad, H. (2015) Time-Aware Authorship Attribution for Short Text Streams. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. N.Y. pp. 727–730.
15. Litvinova, T.A. (2015) Russian written text as a source of the information on the personality of its author (on the material of text corpus of the "personality" new type). *Izvestiya VGPU. Ser.: Pedagogicheskie nauki; Gumanitarnye nauki – Izvestia VGPU. Pedagogical Sciences, The Humanities*. 266:1. pp. 196–198. (In Russian).

16. Porshneva, O.S. (2002) K voprosu ob atributsii tekstov zapisey soldatskikh razgovorov [On the attribution of texts of soldier conversation records]. *Informatsionnyy byulleten' assotsiatsii "Istoriya i komp'yuter"*. 30. pp. 31–44.
17. Kjetsaa, G. (1989) *Kto napisal "Tikhiy Don"?* [The Authorship of the Quiet Don]. Translated from English. Moscow: Kniga.
18. Drozdova, T.N. (2010) Diagnosticheskie i klassifikatsionnye zadachi v avtorovedcheskoy ekspertize blogov [Diagnostic and classification problems in authorship examination of blogs]. *Aktual'nye problemy rossiyskogo prava*. 2 (15). pp. 394–404.
19. Romanov, A.S. (2010) *Metodika i programmnyy kompleks dlya identifikatsii avtora neizvestnogo teksta* [Methods and software system for the identification of the author of an unknown text]. Abstract of Engineering Cand. Diss. Tomsk.
20. Mamaev, M.M. (2015) Gender attribution of translated texts as a specific case of study of an author's linguistic consciousness. *Vestnik MGOU. Ser.: Lingvistika – Bulletin MGOU. Series "Linguistics"*. 2. pp. 25–31.
21. Mukherjee, A. & Liu, B. (2010) Improving Gender Classification of Blog Authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pp. 32–38.
22. Yan, X. & Yan, L. (2006) Gender Classification of Weblog Authors. *Computational Approaches to Analyzing Weblogs*. AAAI. pp. 18–26.
23. Argamon, Sh. et al. (2010) Gender, Genre, and Writing Style in Formal Written Texts. *Text.* 23:3. pp. 321–346. DOI: https://doi.org/10.1515/text.2003.014
24. Rezanova, Z.I., Romanov, A.S. & Meshcheryakov, R.V. (2013) Tasks of author attribution of text in the aspect of gender (on interdisciplinary interaction of linguistics and computer science). *Vestnik Tomskogo gosudartsvennogo univeristeta – Tomsk State University Journal*. 370. pp. 24–28. (In Russian).
25. Brocardo, M.L. (2014) *Authorship Verification for Short Messages using Stylometry*. [Online] Available from: https://www.deepdyve.com/lp/institute-of-electrical-and-electronics-engineers/authorship-verification-for-short-messages-using-stylometry-JM5XWbkHyN. (Accessed: 7 July 2016).
26. Arroju, M. (2015) Age, Gender and Personality Recognition using Tweets in a Multilingual Setting. *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*. pp. 23–31.
27. Rubinshteyn, S.L. (1989) *Osnovy obshchey psikhologii* [Fundamentals of general psychology]. Vol. 1. Moscow: Pedagogika.
28. Pennebaker, J.W., Mehl, M.R. & Niederhoffer, K.G. (2003) Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*. 54:1. pp. 548–571.
29. Vol'f, E.M. (1974) *Grammatika i semantika mestoimeniy* [Grammar and semantics of pronouns]. Moscow: Nauka.
30. Verhoeven, B.S. et al. (2016) *TWISTY: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling.* LREC 2016, Tenth International Conference on Language Resources and Evaluation. pp. 1632–1637.
31. Baranov, A.N. (200) *Vvedenie v prikladnuyu lingvistiku* [Introduction to applied linguistics]. Moscow: Editorial URSS.

Received: 19 January 2017