ФОРМАЛЬНО-ГРАММАТИЧЕСКИЕ КОРРЕЛЯТЫ ПОЛОВОЙ ПРИНАДЛЕЖНОСТИ АВТОРА ПИСЬМЕННОГО ТЕКСТА*

* Работа выполнена при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых: МК-9349.2016.6 - "Языковые средства репрезентации идентичности в автодокументальных текстах: лингвокогнитивное моделирование".

FORMALLY GRAMMATICAL CORRELATES OF THE SEX OF THE AUTHOR OF WRITTEN TEXT

D. Minets
A. Gorushkina

Annotation

The article represents the first stage of the study of the relationship between the formal grammatical parameters of the text and the gender of its author, that is, it reveals one of the aspects of the problem of profiling the personality of the author of a written text. Using the methods of mathematical statistics, a regression model is obtained that links the formal grammatical characteristics of the text (morphological and syntactic) and personality (gender) features of its author.

Keywords: Text, morphological analyzer, formal–grammatical characteristics, linguistic statistics, correlation analysis, gender, gender.

Минец Диана Владимировна
К.филол.н., доцент, Череповецкий государственный университет
Горушкина Анна Валентиновна
Аспирант, Череповецкий государственный университет

Аннотация

Статья представляет первый этап исследования зависимости между формально–грамматическими параметрами текста и половой принадлежностью его автора, то есть раскрывает один из аспектов проблемы профилирования личности автора письменного текста. С применением методов математической статистики получена регрессионная модель, связывающая формально–грамматические характеристики текста (морфологические и синтаксические) и личностные (в частности – половая принадлежность) особенности его автора.

Ключевые слова:

Текст, морфологический анализатор, формально–грамматические характеристики, лингвостатистика, корреляционный анализ, пол, гендер.

На данный момент признается доказанным тезис о зависимости текста как продукта индивидуальной речевой деятельности от личности его автора [8], однако в науке нет единого подхода к методике выявления информации о характеристиках автора текста на основе лингвистического анализа его параметров. В последнее время в связи с интенсивным развитием средств автоматической обработки текста, программ для статистической обработки данных представляется особенно перспективным стилеметрический подход к моделированию личности по тексту, основная идея которого состоит в том, что на большом корпусном материале с использованием методов статистической обработки данных вычисляются корреляции между подлежащими количественной оценке параметрами текстов и характеристиками их авторов [9; 3; 4; 5].

Для решения задач подобного рода в рамках заявленного подхода необходимы корпус текстов, специально созданный для решения данной задачи и содержащий метаразметку; перечень параметров текста, являющихся информативными для диагностирования той или иной характеристики автора (представлен ниже); математические методы выявления корреляций численных значений параметров текстов и характеристик личности их авторов.

В рамках настоящей статьи представлены результаты проведенного авторским коллективом эксперимента по выявлению зависимости между формально–грамматическими, поддающимися квантификации параметрами текста и полом автора на материале специального созданного корпуса автодокументальных текстов с применением статистических методов обработки данных. Автодокументальный (мемуарно–автобиографический) дискурс – активная среда реализации специфических параметров категории идентичности, вербализующихся на разных уровнях языковой структуры [7, с. 66]. Жанровые модификации автодокументалистики – дневник, мемуары, автобиография, записки, сводные тетради – специфическая сфера отражения концептов идентичности.

Общий корпус авторских текстов для анализа представлен текстовыми фрагментами (средний объем 480 – 520 лексических единиц) произведений различных жанровых модификаций мужской и женской автодокументальной прозы XVIII – XX вв. Корпус насчитывал свыше 1000 фрагментов 150 авторов (мужчин и женщин).

Методами автоматической обработки текстов с использованием специально разработанного на основе на основе морфологического анализатора mystem от Яндекс [10] были извлечены числовые значения формально-грамматических параметров текста, список которых был составлен по материалам русскоязычной и англоязычной научной литературы и резюмирован рядом работ подобного профиля [2; 6].

Для построения модели зависимости пола от морфолого-синтаксических параметров текста использовался линейный дискриминантный анализ Фишера. При этом методе записывается функция линейной регрессии

$$Y = a + b_1X_1 + b_2X_2 + ... + b_mX_m \ ,$$

где $X_1, X_2, ..., X_m$ – значения факторов, описывающих объект и влияющих на результат,

Y – результирующая переменная.

В дискриминантном анализе значение результирующей переменной сравнивается с пороговым значением p. В случае, если $Y < p$, объект относится к первой группе, в противном случае – ко второй.

В нашем случае факторы – характеристики текста, предложенные Т.А. Литвиновой [6, с. 133], а также параметры, выделенные И.Н. Кавинкиной [2] по методике Е.И. Горошко [1, с. 63] – коэффициент предметности, равный отношению количества существительных и местоимений к количеству прилагательных и глаголов, коэффициент качественности, равный отношению числа прилагательных и наречий к количеству существительных и глаголов, коэффициент активности, равный количеству глаголов, причастий и деепричастий к общему количеству слов в тексте, коэффициент динамизма, равный количеству глаголов, причастий и деепричастий к количеству существительных, прилагательных и местоимений.

Значимым является также анализ синтаксического уровня текстов, в частности, структуры предложений, однако в настоящее время он недостаточно поддается автоматизации, в связи с чем было взято ограниченное число параметров текстов на синтаксическом уровне: количество предложений; количество сложных предложений; количество сложных бессоюзных предложений. Все эти параметры также оказываются значимыми для моделирования пола автора по тексту.

Полный же список
рассматриваемых характеристик текста - следующий:

Кол-во имен существительных	x_1
Кол-во глаголов	x_2
Кол-во личных местоимений	x_3
Кол-во указательных местоимений	x_4
Кол-во относительных и вопросительных местоимений	x_5
Общее количество местоимений всех разрядов	x_6
Кол-во местоименных наречий	x_7
Кол-во деепричастий	x_8
Кол-во причастий	x_9
Кол-во предлогов	x_{10}
Кол-во союзов	x_{11}
Кол-во частиц	x_{12}
Кол-во знаменательных слов	x_{13}
Кол-во незнаменательных слов	x_{14}
Общее кол-во слов в тексте	x_{15}
Кол-во сложных предложений	x_{16}
Кол-во сложных бессоюзных предложений	x_{17}
Кол-во существительных и местоимений	x_{18}
Кол-во прилагательных и глаголов	x_{19}
Кол-во прилагательных и наречий	x_{20}
Кол-во глаголов и существительных	x_{21}
Кол-во глаголов, причастий и деепричастий	x_{22}
Кол-во существительных, прилагательных и местоимений	x_{23}
Кол-во предлогов и союзов	x_{24}

В качестве факторных переменных X_i, кроме перечисленных выше, естественно взять отношения вида x_i/x_{15}, где в знаменателе находится общее количество слов в тексте. Значения этих переменных принадлежат отрезку $[0; 1]$ и выражают долю соответствующих частей речи в тексте. Деление на общее количество слов необходимо, поскольку рассматриваемые тексты имеют разную длину. По аналогичной причине рассматриваются переменные x_{20}/x_{21} и x_{18}/x_{19}.

Для построения дискриминирующей функции необходимо выбрать факторные переменные и определить коэффициенты линейной регрессии b_i и константу a. Коэффициенты регрессии отражают вклад переменных в результат: чем больше коэффициент регрессии, тем сильнее соответствующая переменная влияет на результат. Константа регрессии a используется для цент-

рирования значений регрессионной функции для удобства выбора порогового значения p (в нашем случае естественно выбрать $p = 0.5$).

Наилучшая модель должна учитывать максимальное возможное число факторов, поэтому в качестве множества факторных переменных было выбрано следующее

$$1, \frac{x3}{x15}, \frac{x5}{x15}, \frac{x7}{x15}, \frac{x13}{x15}, \frac{x14}{x15}, \frac{x16}{x15}, \frac{x17}{x15}, \frac{x18}{x15}, \frac{x22}{x15}, \frac{x23}{x15}, \frac{x18}{x19}, \frac{x20}{x21},$$

Переменные x_1, x_2, x_4, x_6, x_8, x_9, x_{10} исключены из рассмотрения по одной из следующих причин:

1. чтобы избежать мультиколлинеарности, при которой одна или несколько переменных линейно выражаются через остальные;
2. влияние этих переменных на результат незначительно (соответствующие коэффициенты регрессии bi близки к нулю).

Для определения коэффициентов регрессии b_i и константы a применяется метод наименьших квадратов. Процедура определения этих величин встроена во многие стандартные математические пакеты. Мы использовали систему компьютерной математики Maple 14, содержащую пакет статистического анализа *Statistics*. Функция *LinearFit* этого пакета вычисляет коэффициенты регрессии. Исходные данные представляли собой 1015 фрагментов текстов различных авторов, среди которых 514 фрагментов за мужским авторством и 501 фрагмент – за женским. С помощью функции *LinearFit* нами получена следующая формула для значения результирующей переменной:

$$\text{"Y"} = \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} + \frac{6,9\,x13}{x15} - 9,7$$

Анализ распределения величины Y показал, что она распределена нормально со средним значением 0.5. Поэтому в качестве порогового значения было выбрано $p = 0.5$. Таким образом, алгоритм определения пола автора текста, следующий: если значение Y меньше 0.5, то пол автора – женский, иначе – мужской. Значения коэффициентов регрессии показывают, что вклад различных переменных вполне сопоставим: отношение наибольшего коэффициента к наименьшему не превосходит 20.

Попытка исключить переменные с наименьшими коэффициентами регрессии увеличивает процент ошибок, хотя и незначительно. Поэтому было решено остановиться на этом варианте формулы.

Оценка точности модели показала следующее. При обучении на полной выборке из 1015 фрагментов текстов процент ошибок составляет 29.8, то есть формула даёт верный результат примерно в 70% случаев.

Кроме того, было рассмотрено несколько случайных разбиений исходной выборки на две примерно равные по объёму части. Одна часть использовалась для обучения, а другая – для проверки точности модели.

Во всех случаях точность модели оказалась не менее 68%. Поскольку приведённая выше формула построена по выборке вдвое большего объёма, то можно рассчитывать на несколько большую точность определения пола.

Таким образом, представляется обоснованным вывод о том, что точность определения пола с помощью полученной модели равна 70%.

Отметим, что настоящее исследование является пилотным и лишь намечает направления поиска в области профилирования пола автора по тексту на основе морфолого-синтаксических параметров, однако даже на данном этапе подтверждена корреляция между морфолого-синтаксическими параметрами текста действительно существует, и исследования в этом направлении должны быть продолжены.

ЛИТЕРАТУРА

1. Горошко, Е.И. Особенности мужского и женского стиля письма // Преображение, 1998. № 6. С. 48–64.
2. Кавинкина, И.Н. Проявление гендера в речевом поведении носителей русского языка: монография. Гродно: ГрГУ, 2006.
3. Литвинова, Т. А. Установление характеристик (профилирование) автора письменного текста // Филологические науки. Вопросы теории и практики. 2012. № 2 (13). С. 90–94.
4. Литвинова, Т. А. Языковые корреляты личностных особенностей автора письменного текста: алгоритм исследования // В мире научных открытий. Серия: Проблемы науки и образования. 2012. № 9.3 (33). С. 236–255.

5. Литвинова, Т. А., Загоровская О. В., Середин П. В., Лантюхова Н. Н., Шевченко И. С. Профилирование автора письменного текста: подходы, методы и их оптимизация // Филология, искусствоведение и культурология: актуальные вопросы и тенденции развития: материалы международной заочной научно–практической конференции (13 мая 2013 г.). Новосибирск: Изд. "СибАК", 2013. С. 69–79.
6. Литвинова, Т.А. Формально–грамматические корреляты личностных особенностей автора письменного текста // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2013. № 12 (30): в 2–х ч. Ч. I. С. 132–135.
7. Минец, Д.В. Практики (авто)биографической идентификации в автодокументальном дискурсе // Череповецкие научные чтения – 2016: Материалы Всероссийской научно–практической конференции: В 3ч. Ч.1. Литературоведение, лингвистика, СМИ, история, философия, социология, политология, художественное образование / Отв. ред. Е.В. Целикова. Череповец: ЧГУ, 2017. С.66–67.
8. Фомина, Н.А. Свойства личности и особенности речевой деятельности. Рязань: Узорочье, 2002.
9. Lyons, J. Linguistic Semantics. Cambridge: Cambridge University Press, 1995.
10. MyStem. Яндекс [Эл.ресурс]. Режим доступа: https://tech.yandex.ru/mystem (дата обращения: 01.07.2017г.).