УДК 81'33

РУССКИЙ ПИСЬМЕННЫЙ ТЕКСТ КАК НОСИТЕЛЬ ИНФОРМАЦИИ ОБ ИНДИВИДУАЛЬНО-ЛИЧНОСТНЫХ ХАРАКТЕРИСТИКАХ ЕГО АВТОРА (НА МАТЕРИАЛЕ КОРПУСА ТЕКСТОВ НОВОГО ТИПА PERSONALITY)

ЛИТВИНОВА Татьяна Александровна,
кандидат филологических наук, научный сотрудник Регионального центра русского языка,
Воронежский государственный педагогический университет

АННОТАЦИЯ. *В статье представлены результаты исследований, выполненных на материале создаваемого автором принципиально нового корпуса текстов Personality, содержащего, помимо образцов естественной письменной речи, данные об их авторах (пол, возраст, результаты психологического тестирования). Исследования с применением методов автоматической обработки языка и вычислительной лингвистики показали наличие статистически достоверных корреляций между формально-грамматическими характеристиками текстов и особенностями личности их авторов.*
КЛЮЧЕВЫЕ СЛОВА: *корпус текстов, корпусная лингвистика, математическая лингвистика, диагностирование личности по тексту, компьютерная лингвистика, авторoведение, авторoведческая экспертиза.*

LITVINOVA T.A.,
Cand. Philol. Sci., Scientific Fellow of the Regional Centre of Russian Language,
Voronezh State Pedagogical University

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)

ABSTRACT. *The articles presents the results of the study of the materials provided by a text corpus Personality designed by the author, which includes along with samples of natural written speech the author details (gender, age, psychological testing results). The study identified statistically valid correlations between the formal grammar parameters of texts and personality traits of their authors.*
KEY WORDS: *text corpus, corpus linguistics, computational linguistics, author profiling, computer linguistics, authorship attribution, forensic authorship attribution.*

В настоящее время является общепризнанным положение о том, что любой текст, в том числе анонимный, несет в себе социобиографическую информацию о его авторе, которая может быть установлена путем специального лингвистического анализа. Данная информация может быть получена даже при ее намеренном искажении, поскольку транслируется на языковых уровнях, неподконтрольных сознанию (например, на грамматическом). Однако специальных исследований по данному вопросу на материале русского языке крайне недостаточно.

В российской науке приоритет в изучении данной проблемы принадлежит психологам, однако, как отмечает К.Ф. Седов, несмотря на то, что взаимосвязь личностных психологических характеристик человека и его дискурсивного поведения для большинства психологов сегодня очевидна с позиции здравого смысла, **последовательного и системного исследования** такой взаимосвязи в социопсихолингвистике еще не проводилось, в этом направлении исследователями делаются лишь первые шаги [2].

В зарубежной науке приоритет в исследовании проблемы диагностирования личности по тексту также принадлежит психологам. Однако уже с 1990-х гг. к решению данной проблемы подключаются лингвисты, математики и информатики, и в данной сфере исследований начинается активное использование методов математической статистики, компьютерной лингвистики, в частности средств автоматической обработки языка, что позволяет быстро анализировать большие массивы текстового материала. На основе найденных корреляций между численными значениями поддающихся квантификации параметров текста и баллами по шкалам психотестов, полученными авторами текстов, исследователями строятся математические модели и разрабатываются программные средства для автоматизированного диагностирования характеристик личности по тексту (в виде баллов по шкалам тестов).

Применительно к русскому языку до настоящего времени отсутствовали работы по диагностированию

Информация для связи с автором: centr_rus_yaz@mail.ru

индивидуально-психологических особенностей автора текста на основе анализа численных значений формально-лингвистических параметров текста. Наши работы [3; 4; 5], направленные на построение математических моделей для диагностирования индивидуально-психологических характеристик автора письменного текста на основе численных значений формально-лингвистических параметров текста, на материале специально созданного корпуса текстов корпус текстов разных жанров, представляющих образцы естественной письменной речи (описание картины, эссе на различные темы и пр.) и снабженных информацией об их авторах (пол, возраст, данные психологического тестирования) с применением методов автоматической обработки языка, показали наличие устойчивых корреляций между некоторыми формально-грамматическими параметрами текста и характеристиками личности. Материалом для исследования послужил корпус текстов Personality [6] – создаваемый под руководством автора корпус текстов разных жанров, представляющих образцы естественной письменной речи. В настоящее время в корпусе представлены тексты более 700 респондентов, и корпус постоянно пополняется.

Для данного этапа исследований нами были проанализированы 200 текстов от 200 респондентов. Все тексты были размечены при помощи свободно распространяемого морфологического парсера фирмы «Xerox» (https://open.xerox.com/Services/fst-nlp-tools/Consume/Part%20of%20Speech%20Tagging%20(Standard)-178), который был использован нами ранее и показал высокую точность, далее было произведено извлечение числовых значений выбранных параметров текста. Данные для расчетов были занесены в Excel. Далее данные были экспортированы в программу SPSS Statistics и произведен корреляционный анализ между числовыми значениями выбранных параметров текста и баллами по шкалам теста (отдельно для каждой шкалы), $p < 0,05$. Далее на основе найденных корреляций были построены уравнения регрессии – математические модели для диагностирования пола авторов текстов и некоторых психологических характеристик – и произведена оценка эффективности этих моделей на независимой выборке. В качестве языковых параметров использовались формально-грамматические характеристики текста [3], доказавшие свою эффективность при решении задачи диагностирования личности автора текста в исследованиях на материале английского языка [7], отдельное исследование было проведено для выявления корреляций между характеристиками личности автора текста и частотностями биграмм частей речи [4].

В среднем точность полученных моделей для диагностирования индивидуально-психологических характеристик авторов (определенных при помощи психологического теста «Большая пятерка») составила 60–65%, что сопоставимо с результатами исследований на материале английского языка. Насколько нам известно, это первый в российской лингвистике опыт построения комплексных прогностических моделей, учитывающих сразу несколько параметров письменного текста и применимых к решению задачи прогнозирования пола и некоторых психологических характеристик автора конкретного письменного текста.

Таким образом, применив распространенный в современной зарубежной науке подход к решению задачи моделирования личности по тексту, предполагающий использование специально составленного корпуса текстов, содержащих метаданные в виде информации об их авторах (пол, возраст, результаты психологического тестирования в виде баллов), анализ преимущественно формально-грамматических характеристик текста, поддающихся квантификации и извлечению современными средствами автоматической обработки языка, а также построение прогностических математических моделей на основе выявленных корреляций между числовыми значениями параметров текста и баллами по шкалам психологических тестов к русскому языку, мы получили математические модели для диагностирования индивидуально-психологических характеристик автора текста, точность которых сравнима с точностью подобных моделей, построенных для английского языка. Однако данный подход, при всей своей эффективности, имеет недостатки, о которых говорят и авторы работ, выполненных на материале английского языка: поскольку параметры текстов выбираются без опоры на какую-либо теорию, полученные корреляции между формально-грамматическими параметрами текста и характеристиками личности не находят своего объяснения. Кроме того, параметры в основном отражают особенности речевого произведения на уровне морфологии, частично – синтаксиса (на уровне предложения); характеристики же, присущие только тексту (например, параметры, отражающие особенности употребления средств связи между предложениями), не анализируются, так как слабо поддаются автоматизированному подсчету.

Как представляется, для разработки более эффективных методик диагностирования индивидуально-психологических характеристик личности по тексту необходим синтез имеющихся достижений в этой области (создание специальных корпусов текстов, содержащих метаданные в виде информации об их авторах, применение средств автоматической обработки языка для лингвистической разметки корпусов, применение программных средств для автоматического извлечения числовых значений выбранных параметров текстов; использовании современных математических пакетов для методов обработки данных и построения прогностических математических моделей) и принципиально новый подход к выбору параметров текста, которые могут коррелировать с теми или иными индивидуально-психологическими характеристиками личности. На наш взгляд, комплексный многоуровневый анализ текста с привлечением данных психологии, в том числе такого ее направления, как нейропсихология индивидуальных различий, а также данных психолингвистики, нейролингвистики позволит более системно и полно описать взаимосвязь индивидуально-психологических особенностей личности и характеристик ее речевой продукции, а также дать возможное объяснение найденным корреляциям.

СПИСОК ЛИТЕРАТУРЫ:

1. Доброва В.В. Взаимосвязь монорематических высказываний и личностных характеристик субъекта диалога : автореф. дис. ... канд. псих. наук / В.В. Доброва. – Самара, 2008. – 199 с.
2. Седов К.Ф. Дискурс и личность : эволюция коммуникативной компетенции / К.Ф. Седов. – М. : Лабиринт, 2004. – 320 с.
3. Литвинова Т.А. Формально-грамматические корреляты личностных особенностей автора письменного текста / Т.А. Литвинова // Филологические науки. Вопросы теории и практики. – 2013. – № 12(30). – Ч. 1. – С. 132–135.
4. Литвинова Т.А. Частоты встречаемости последовательностей частей речи в тексте и психофизиологические характеристики его автора: корпусное исследование / Т.А. Литвинова, О.А. Литвинова, П.В. Середин // Вестник Иркутского государственного лингвистического университета. – 2014. – № 2. – С. 9–13.
5. Litvinova T.A. Profiling the author of a written text in Russian / T.A. Litvinova // Journal of Language and Literature. – 2014. – № 5(4). – P. 210–216.
6. Загоровская О.В. Электронный корпус студенческих эссе на русском языке и его возможности для современных гуманитарных исследований / О.В. Загоровская, Т.А. Литвинова, О.А. Литвинова // Мир науки, культуры и образования. – 2012. – № 3(34). – С. 387–389.
7. Литвинова Т.А. Языковые корреляты личностных особенностей автора письменного текста: алгоритм исследования / Т.А. Литвинова // В мире научных открытий. Серия: Проблемы науки и образования. – 2012. – № 9.3(33). – С. 236–255.