УДК 81.33

ВОЗМОЖНОСТИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ДИАГНОСТИРОВАНИИ ЛИЧНОСТИ ПО ТЕКСТУ[1]

Т. А. Литвинова

Региональный центр русского языка
Воронежского государственного педагогического университета

Поступила в редакцию 27 марта 2015 г.

Аннотация: *проблема извлечения из текста информации о личности его автора путем лингвистического анализа имеет как теоретическую, так и практическую значимость и активно разрабатывается в современной науке. Для решения данной задачи необходимы корпуса текстов, содержащие метаразметку в виде данных об их авторах, список параметров текста, поддающихся квантификации, и соответствующий математический аппарат. Важное значение для решения данной задачи имеетиспользование средств автоматической обработки языка, которые позволяют упростить процесс извлечения численных значений параметров текста. В статье представлены результаты исследований автора, выполненных на материале создаваемого автором принципиально нового корпуса текстов «Personality».Описываются выявленные корреляции между формально-грамматическими параметрами текстов и характеристиками личности их авторов (пол, психологические особенности).*
Ключевые слова: *корпус текстов, диагностирование личности по тексту, компьютерная лингвистика, авторoведение, авторoведческая экспертиза.*

Abstract: *the problem of extracting personality data from the text by means of the linguistic analysis is theoretically and practically significant and is a widely discussed issue in modern linguistics. To deal with this problem, researchers need specialized text corpora containing metatags including the author details, a list of text parameters to be quantified and a relevant mathematical apparatus. The use of automated language processing tools for facilitating the extraction of numerical text parameters is also of major importance. The article deals with the results of the author's research based on anewly developed «Personality» text corpus. The correlations between the formal grammatical parameters of text and characteristics of authors' personalities (e.g. gender, psychological traits) are described.*
Key words: *text corpus, author profiling, computer linguistics, authorship attribution, forensic authorship attribution.*

В мировой науке активно развиваются исследования, направленные на разработку методик, автоматически извлекающих из текста различную информацию о его авторе (пол, возраст, психологические характеристики, уровень образования и др.) на основе анализа его формально-грамматических характеристик [1].

Основной подход к решению данной задачи состоит в создании специальных корпусов текстов, содержащих, помимо собственно текстов, метаразметку в виде информации о личности их авторов (пол, возраст, баллы по шкалам психологических тестов и пр.), разметке корпусов средствами автоматической обработки языка, извлечении численных значений, поддающихся квантификации параметров текста, вычислении корреляций между этими значениями и характеристиками личности автора текста и построении на их основе математических моделей для диагностирования тех или иных характеристик автора текста [2].

Решение задач диагностирования личности автора письменного текста (*author profiling*) при помощи математических методов и средств автоматической обработки языка имеет большое значение не только теоретическое, но и практическое, в частности для криминалистических экспертиз, исследований рынка, вследствие чего данная проблематика активно разрабатывается в мировой науке в последние годы. О большом интересе к указанной проблеме свидетельствует, к примеру, тот факт, что с 2013 г. задача диагностирования личности по тексту ставится перед участниками ежегодного международного конкурса PAN, посвященного проблемам выявления плагиата и установлению авторства, в том числе текстов интернет-коммуникации [3]. Каждый год организаторы предоставляют участникам корпуса текстов и ставят

[1] Исследование выполнено при поддержке гранта РФФИ № 13-06-00016 «Моделирование личности автора письменного текста», гранта РГНФ № 13-14-36001 «Речевой портрет воронежских студентов (на материале электронного корпуса текстов «Россия и мир глазами воронежских студентов)».

те или иные задачи по выявлению информации о личности их авторов. Те команды, чьи методики дают наиболее точный результат, становятся победителями. Так, в 2013 г. перед участниками ставилась задача диагностирования пола и возраста авторов текстов в соцсетях, на английском и испанском языках [4]. Данная задача привлекла наибольшее число участников: 21 команда пыталась с помощью специально разработанных методик выявить пол и возраст авторов текстов, анализируя как лексический уровень текста, так и уровень формально-грамматический. Также перед участниками ставилась задача выявления психологических характеристик и родного языка авторов текстов. В 2014 г. организаторы конкурса также ставили перед участниками задачи по диагностированию личности авторов текстов в социальных медиа, а также по установлению авторства текстов интернет-коммуникации [5]. Объявленные на 2015 г. задачи включают в себя диагностирование пола, возраста, психологических характеристик авторов твитов на английском, испанском, итальянском, датском языках [6].

Заметим, что большинство научных работ, посвященных разработке методик диагностирования личности автора текста, выполнено на материале английского языка, но анализ работ, представленных на PAN, свидетельствует о том, что с каждым годом число языков, на материале которых проводятся данные исследования, расширяется.

Насколько нам известно, применительно к русскому языку до настоящего времени отсутствовали работы по диагностированию индивидуально-психологических особенностей автора текста на основе анализа численных значений формально-лингвистических параметров текста. Работы автора [7–9], направленные на построение математических моделей для диагностирования индивидуально-психологических характеристик автора письменного текста, 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*** [10] – создаваемый под руководством автора 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, профессия, данные психологического тестирования и др.). В настоящее время в корпусе представлены тексты более 1000 респондентов, и корпус постоянно пополняется.

Вкратце опишем полученные результаты. Нами было проанализировано 200 текстов от 100 респондентов, средняя длина текстов одного автора – 166 слов. Параметры автора: пол, характеристики личности, оцененные по 5-факторному личностному опроснику, более известному как «Большая пятерка» («Великолепная пятерка»). Эта методика была выбрана нами потому, что она является одной из наиболее популярных в зарубежных исследованиях по диагностированию личности по тексту. Именно характеристики личности автора текста, входящие в «большую пятерку», чаще всего и пытаются диагностировать зарубежные исследователи. Нами была использована русскоязычная версия опросника в модификации А. Б. Хромова, позволяющая измерять степень выраженности каждого из пяти факторов (экстраверсия – интроверсия; привязанность – обособленность; самоконтроль – импульсивность; эмоциональная неустойчивость – эмоциональная устойчивость; экспрессивность – практичность) [11].

Также респонденты были протестированы по методике диагностики коммуникативной установки В. В. Бойко [12], измеряющей степень положительности/негативности коммуникативной установки испытуемого.

В качестве параметров текста нами были выбраны различные индексы, отражающие морфологические и синтаксические характеристики текста, всего 67 индексов.

Все тексты были размечены при помощи морфологического парсера фирмы *Xerox* [13]. Синтаксическая разметка проводилась вручную. Далее было произведено извлечение числовых значений выбранных параметров текста. Данные для расчетов были занесены в *Excel*. Затем данные были экспортированы в программу ***SPSS Statistics***, и произведен корреляционный анализ между числовыми значениями параметров текста и характеристиками личности (пол; баллы по тесту «Большая пятерка», отдельно для каждой шкалы), $p < 0.05$ (табл. 1–7). После на основе найденных корреляций были построены уравнения регрессии – математические модели для диагностирования пола авторов текстов и психологических характеристик, и произведена оценка эффективности этих моделей на независимой выборке.

Как видно из табл. 1, для диагностирования пола наиболее значимыми являются параметры текста, описывающие различные соотношения дейктических элементов к общему числу слов, доля в тексте существительных, а также бессоюзных предложений.

Для диагностирования коммуникативной установки автора текста имеет значение доля имен собственных в тексте и сложноподчиненных предложений.

Как видно из таблиц, для диагностирования психологических характеристик автора текста имеют значение как морфологические, так и синтаксические характеристики текста.

Отдельное исследование было проведено для выявления корреляций между характеристиками

Таблица 1

Пол									
Параметры текста	Знаменательных слов / незнаменательных слов	Существительных / всего слов	Незнаменательных словоупотреблений / существительных	(Указат. мест. + вопросит.-относит. мест.+ + личных мест. + мест.-нар.) / всего слов	(Местоим. всех разрядов + местоим. нар.) / всего слов	Бессоюзных сложных предложений / сложных предложений всего	(Местоим. всех разрядов + союзы + частицы) / всего слов	(Мест. + частиц + союзов) / (сущ. + нар. + прил. + глаг. + междомет. + деепр. + прич.)	Личных местоимений / всего слов
Коэффициент корреляции	0.258	0.252	–0.297	–0.325	–0.269	0.253	–0.286	–0.272	–0.274

Таблица 2

Коммуникативная установка			
Параметры текста	Всего сложноподчиненных предложений / сложных предложений	Имена собственные / всего слов	Имена собственные / (всего сущ. + личн. мест.)
Коэффициент корреляции	–0.255	0.341	0.339

Таблица 3

Экстраверсия / интроверсия							
Параметры текста	Простых предложений / предложений всего	Причастий + деепричастий / всего слов	Союзов / предлогов	Указ. мест. + вопросит.-относит. мест. / всего слов	Предлогов / всего слов	Дееприч. оборотов + прич. оборотов / всего обособлений	Деепричастий / всего слов
Коэффициент корреляции	0.232	–0.245	0.257	0.33	–0.232	–0.236	–0.351

Таблица 4

Привязанность / обособленность						
Параметры текста	Всего слов / всего простых предложений	Предлогов / всего незнаменательных слов	(Мест. + предлогов) / всего незнаменательных слов	Союзов / предлогов	Предлогов / слов всего	Деепричастий / всего слов
Коэффициент корреляции	–0.246	–0.257	–0.23	0.276	–0.267	–0.347

Таблица 5

Самоконтроль / импульсивность				
Параметры текста	Прилагательных/наречий	Причастий + деепричастий / всего слов	Указ. мест. + относит.-вопросит. мест. / всего слов	Деепричастий / всего слов
Коэффициент корреляции	–0.267	–0.242	0.233	–0.329

Таблица 6

Эмоциональная устойчивость / эмоциональная неустойчивость		
Параметры текста	Прил. / наречий	Деепричастий / число слов
Коэффициент корреляции	–0.287	–0.272

Таблица 7

Экспрессивность / практичность							
Параметры текста	Союзов / незнаменательных слов	Частиц / всего незнаменательных употреблений	Причастий + деепричастий / всего слов	Указ. мест. + относит.-вопросит. мест. / всего слов	Существительных / местоимений	Частиц / всего слов	Деепричастий / всего слов
Коэффициент корреляции	0.237	–0.285	–0.33	0.268	–0.25	–0.294	–0.417

личности автора текста и частотностями биграмм (последовательностей из двух элементов) частей речи [4].

Методами автоматической обработки языка (также использовался морфологический анализатор фирмы *Xerox*) для каждого текста были рассчитаны частоты встречаемости биграмм частей речи (всего в анализируемом нами материале было зафиксировано 227 типов биграмм), затем были выбраны биграммы, встречающиеся не менее чем в 75 % проанализированных текстов (табл. 8).

Далее были вычислены доли каждой из биграмм в текстах (число биграмм каждого типа делили на общее число слов в тексте), после чего традиционными математическими методами установлены корреляции между каждым из параметров текста, в роли которых выступали доли в тексте самых частотных биграмм, и характеристикой личности, приведенной к числовому значению (пол, принимали для расчетов: женщина – 0, мужчина – 1; баллы по 5 шкалам теста) (табл. 9).

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.

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 основы 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 и принципиально новый подход к выбору параметров текста, которые могут коррелировать с теми или иными индивидуально-психологическими характеристиками личности. Нами был применен подход к выбору параметров текста с привлечением данных психологии, в том числе такого ее направления, как нейропсихология индивидуальных различий, а также данных психолингвистики, нейролингвистики, и в ходе пилотного эксперимента был получен ряд корреляций между психологическими характеристиками личности, связанными с риском аутоагрессивного поведения, и формально-лингвистическими параметрами текста (индексы удобочитаемости текста, индексы лексического разнообразия, морфолого-синтаксические параметры и др.), и предпринята попытка объяснения полученных корреляций с точки зрения данных нейронаук [5].

По нашему мнению, подход к выбору параметру текстов для диагностирования личности их авторов

Таблица 8

adj-noun	cm-conj	conj-noun	det-noun	noun-cm	noun-conj	noun-noun	noun-prep	noun-sent	noun-vfin	pers-vfin	pers pers	prep-adj	prep-noun	prep pers	ptcl vfin	sent-noun	sent pers	vfin-cm	vfin vfin	vfin-noun	vfin-prep

Таблица 9

Характеристика автора	Биграммы	Коэффициент корреляции
Пол	prep_noun	0.215
Пятифакторная модель		
Экстраверсия/интроверсия	pers-vfin	0.304
Привязанность/обособленность	pers-vfin ptcl-vfin	0.297 0.321
Эмоциональная неустойчивость / эмоциональная устойчивость	adj-noun noun-prep prep-noun	–0.405 –0.414 –0.322
Экспрессивность/практичность	noun-prep	–0.506

с учетом данных нейролингвистики, нейропсихологии имеет большие перспективы, так как, во-первых, полученные с его помощью результаты внесут свой вклад в решение проблемы взаимосвязи языка и личности, а во-вторых, увеличат точность прогностических моделей.

ЛИТЕРАТУРА

1. *Argamon Sh.* Automatically Profiling the Author of an Anonymous Text / Sh. Argamon, M. Koppel, J. W. Pennebaker, J. Schler // ACM. – 2009. – № 52 (2). – P. 119–123.

2. *Литвинова Т. А.* Языковые корреляты личностных особенностей автора письменного текста: алгоритм исследования / Т. А. Литвинова // В мире научных открытий. Сер. : Проблемы науки и образования. – 2012. — № 9.3 (33). – С. 236–255.

3. *Rosso P.* Uncovering Plagiarism – Author Profiling at PAN / P. Rosso, F. Rangel. – Mode of access: http://ercim-news.ercim.eu/en96/ri/uncovering-plagiarism-author-profiling-at-pan

4. *Rangel F.* Overview of the Author Profiling Task at PAN 2013 / F. Rangel [et al.] // Working Notes Papers of the CLEF 2013 Evaluation Labs / P. Forner, R. Navigli, D. Tufis (eds). – Mode of access: http://www.clef-initiative.eu/documents/71612/2e4a4d3a-bae2-47f9-ba3c-552ec66b3e04

5. *Rangel F.* Overview of the 2nd Author Profiling Task at PAN 2014 / F. Rangel [et al.]. – Mode of access: www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-papers-final/pan14-author-profiling/rangel14-overview.pdf

6. PAN 2015. – Mode of access: http://pan.webis.de/

7. *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.

8. *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 Иркутск. гос. лингв. ун-та. – 2014. – № 2. – С. 9–13.

9. *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.

10. *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.

11. *Хромов А. Б.* Пятифакторный опросник личности : учеб.-метод. пособие / А. Б. Хромов. – Курган : Изд-во Курган. гос. ун-та, 2000. – 23 с.

12. *Райгородский Д. Я.* Практическая психодиагностика. Методики и тесты : учеб. пособие / Д. Я. Райгородский. – Самара : БАХРАХ, 1998. – 672 с.

13. Xerox morphological analyzer. – Mode of access: https://open.xerox.com/Services/fst-nlp-tools/Consume/Part%20of%20Speech%20Tagging%20(Standard)-178

Региональный центр русского языка Воронежского государственного педагогического университета

Литвинова Т. А., кандидат филологических наук, научный сотрудник

E-mail: centr_rus_yaz@mail.ru

Тел.: 8-980-342-00-73

Regional Centre of Russian Language of the Voronezh State Pedagogical University

Litvinova T. A., Candidate of Philology, Research Assistant

E-mail: centr_rus_yaz@mail.ru

Tel.: 8-980-342-00-73