Особенности исследования личности посредством текста и языковой обработки

Аннотация: теоретическая и практическая значимость получения личных сведений о человеке из текста посредством лингвистического анализа безгранична для науки. Но для проведения такого исследования необходим ряд аспектов: корпус текста с меторазметкой по данным рассматриваемой личности, некоторые параметры текста для квантификации, математическая аппаратура и автоматическая языковая обработка для извлечения из текста числовых значений. В основе настоящей статьи лежат сведения из текстового корпуса «Personality», корреляционные связи формально-грамматических значений и половых и психологических личностных особенностей человека.

Ключевые слова: текстовый корпус, личностный анализ, языковая обработка, лингвистический анализ, информация об авторе.

В настоящее время разрабатываются технологии исследования личности по тексту с использованием компьютерной лингвистики. С их помощью можно определить пол, возраст, психологические характеристики и другие параметры автора текста автоматически, без непосредственного участия человека. При этом предполагается использовать лишь формально-грамматические характеристики текста[1].

Суть данного решения состоит в следующих аспектах:

1) создании корпусов текстов, в которых будет находиться сам текст и его метаразметки, т.е. информация о авторе (половая и возрастная принадлежность, психологические оценки и пр.);
2) автоматической обработке разметок;
3) получение количественных параметров для их последующей квантификации;
4) расчет корреляций полученных данных;
5) выстраивание математической модели с целью последующей диагностики автора[2].

Исследование личности по тексту с использованием компьютерной лингвистики имеет как теоретическую, так и практическую ценность. Что касается практики, то оно будет полезно для проведения криминалистических экспертиз и изучения рынка. Именно поэтому и растет интерес к этим технологиям. И это – не просто слова. Подтверждением является международный конкурс PAN, организованный для расширения опыта обнаружения плагиата и установления авторства с наиболее точными показателями, в том числе в текстах на интернет-порталах.

Так, в 2013 году было дано задание, согласно которому участники определяли половую и возрастную принадлежность автора текста социальной сети, написанного на английском и испанском языках, его психологические особенности и родной язык. Кроме формально-грамматических характеристик анализировались и лингвистические[3]. В 2014 году в задание добавили уже не только соц. сети, но и разные средства интернет-коммуникаций[4]. В 2015 году для исследования было добавлено еще два языка – итальянский и датский[5].

Английский язык в компьютерной лингвистике все же является преобладающим среди научных работ по исследованию личности посредством текста. Однако, как показывает конкурс PAN, языковой аспект становится шире. Первые подобные работы в нашей стране провела Литвинова Т.А., благодаря которым стало понятно, что между

---

[1] Argamon Sh. Automatically Profi ling 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] PAN 2013. – Mode of access: http://pan.webis.de/clef13/pan13-web/index.html

[4] PAN 2014. – Mode of access: http://pan.webis.de/clef14/pan14-web/index.html

[5] PAN 2015. – Mode of access: http://pan.webis.de/clef15/pan15-web/index.html

формально-грамматическими и личностными характеристиками действительно есть устойчивые корреляции[6].

Корпус текстов Personality выступал основой для настоящего исследования[7]. Сегодня он состоит из материалов более 1000 человек, продолжая наполняться. Кратко характеризуя проделанную работу, можно выделить несколько фактов:

- были взяты 200 текстов (средняя длина одного – 166 слов) 100 человек;

- личностные характеристики оценены по системе «Большой пятерки», так как она пользуется наибольшей популярностью у иностранных специалистов;

- применена версия опросника А.Б. Хромова, выражающая преобладание экстраверсии-интроверсии; привязанности-обособленности; самоконтроля-импульсивности; эмоциональной неустойчивости- устойчивости; экспрессивности-практичности)[8];

- методика В.В. Бойко позволила осуществить проверку коммуникативной настройки.

Из параметров текста было отдано предпочтение 67-и индексам, показывающим его морфологические и синтаксические характеристики. Морфологическая разметка осуществлялась на парсере компании Xerox, синтаксическая – вручную[9]. После этого были рассчитаны числовые показатели, которые поместили в таблицу Excel, а из нее перенесли в SPSS Statistics. Далее – рассчитана корреляция ($p < 0,05$) между числовыми показателями текста и личностными характеристиками, при этом (табл. 1-7), а в конце – построены уравнения регрессии, которые представляют собой математические модели, и оценена их эффективность на случайной выборке.

Для определения пола важны критерии текста, которые показывают отношение дейктических элементов к общему количеству слов, процент существительных и бессоюзных предложений (табл. 1). В свою очередь, для определения коммуникативной настройки важны имена собственные и некоторое количество сложносочиненных предложений. Для психологических характеристик автора морфологические и синтаксические характеристики текста.

Таблица 1 – Диагностирование пола

Параметры текста				
Показатель	**Знаменательных / незнаменательных слов**	**Существительных / всего слов**	**Указат. мест + вопросит.-относит. мест. + личных мест. + мест.-нар. / всего слов**	**Местоим. всех разрядов + местоим. нар. / всего слов**
Коэффициент корреляции	0,258	0,252	-0,297	-0,325
	Бессознательных сложных предложений / сложных	**Местоим. всех разрядов + союзы + частицы / всего слов**	**Мест. + частиц + союзов / сущ. + нар. + прил. + глаг. + междомет. + деепр. + прич.**	**Личных местоимений / всего слов**

[6] 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; 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; 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.

[7] 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.

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

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

	предложений всего			
	0,253	-0,286	-0,272	-0,274

Таблица 2 – Диагностирование коммуникативной установки

Параметры текста			
Показатель	**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 – Диагностирование экстраверсии/интроверсии

Параметры текста							
Показатель	**Простых предлож-й / предлож-й всего**	**Прич. + деепр. / всего слов**	**Союзов / предлогов**	**Указ. мест. + вопросит. -относит. мест. / всего слов**	**Предлогов / всего слов**	**Деепр. оборот. + прич. оборот. / всего обособ-ий**	**Деепр. / всего слов**
Коэффициент корреляции	0,232	-0,245	0,257	0,33	-0,232	-0,236	-0,351

Таблица 4 – Диагностирование привязанности/обособленности

Показатель	**Всего слов / всего простых предлож-ий**	**Предлогов / всего незнаменат-ых слов**	**Мест. + предлогов / всего незнаменат-ых слов**	**Союзов / предлогов**	**Предлогов / всего слов**	**Деепр. / всего слов**
Коэффициент корреляции	-0,246	-0,257	-0,23	0,276	-0,267	-0,347

Таблица 5 – Диагностирование самоконтроля/импульсивности

Параметры текста				
Показатель	**Прил. / нар.**	**Прич. + деепр. / всего слов**	**Указ. мест. + вопросит.-относит. мест. / всего слов**	**Деепр. / всего слов**
Коэффициент корреляции	-0,267	-0,242	0,233	-0,329

Таблица 6 – Диагностирование эмоциональной устойчивости и неустойчивости

Показатель	**Прил. / нар.**	**Деепр. / всего слов**
Коэффициент корреляции	-0,287	-0,272

Таблица 7 – Диагностирование экспрессивности/практичности

Параметры текста							
Показатель	**Союзов / незнаменат-ых слов**	**Частиц / всего незнаменат-ых употреб-ий**	**Прич. + деепр. / всего слов**	**Указ. мест. + вопросит.-относит. мест. / всего слов**	**Сущ. / мест.**	**Частиц / всего слов**	**Деепр. / всего слов**
Коэффициент корреляции	0,237	-0,285	-0,33	0,268	-0,25	-0,294	-0,417

Отдельное исследование было посвящено поиску устойчивых корреляций между личностными характеристиками автора текста и частотностями биграмм частей речи[10]. Автоматической обработкой языка и тем же парсером от Xerox каждому тексту была

[10] PAN 2013. – Mode of access: http://pan.webis.de/clef13/pan13-web/index.html

высчитана регулярность повторов биграмм (зафиксировано 227 типов) частей речи. После этого отобраны такие вариации, которые встречались минимум в 75% случаях и выведены доли каждой и найдены устойчивые корреляции (табл. 8).

Таблица 8 – Корреляция биграмм и личностных характеристик

Личностная характеристика	**Биграмма**	**Коэффициент корреляции**
Пол	prep_noun	0,215
Пятифакторная модель		
Экстраверсия/интроверсия	pers-vfi n	0,304
Привязанность/обособленность	pers-vfi n ptcl-vfi n	0,297 0,321
Эмоциональная неустойчивость / устойчивость	adj-noun noun-prep prep-noun	-0,405 -0,414 -0,322
Экспрессивность/практичность	noun-prep	-0,506

Точность математических моделей для исследования личности по тексту составила от 60 до 65% стала первым исследованием в российской лингвистике, не сильно отличающаяся от результатов аналогичного диагностирования на базе английского языка. Несмотря на то, что такой комплексный подход является эффективным, у него есть и недостатки, о которых, к слову, упоминается и в научных работах на английском языке:

1) все устойчивые корреляции между формально-грамматическими характеристиками текста и личностными характеристиками автора необъяснимы из-за отсутствия теории;

2) отражается специфика речевого произведения на уровне морфологии и частично на уровне синтаксиса, а вот параметры, которые присущи только тексту, не анализируются.

Для эффективности необходимо синтезировать достижения и выбирать параметры, которые могут иметь устойчивые корреляции с психологическими характеристиками автора. В настоящем исследовании применен подход на основе нейропсихологических различий, психо- и нейролингвистики. Вычислены корреляции между личностными характеристиками психологического начала, имеющие аутоагрессивное поведение, и формально-лингвистическими составляющими текста: индексами удобочитаемости, лексического разнообразия, морфологосинтаксических параметров и пр. Предпринята попытка объяснить результат с точки зрения нейронаук. Такой подход вносит вклад в установление взаимосвязи между языком и личностью, способствуя увеличению точности прогностических моделей.

Литература

1. Argamon Sh. Automatically Profi ling 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. PAN 2013. – Mode of access: http://pan.webis.de/clef13/pan13-web/index.html
4. PAN 2014. – Mode of access: http://pan.webis.de/clef14/pan14-web/index.html
5. PAN 2015. – Mode of access: http://pan.webis.de/clef15/pan15-web/index.html
6. 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.
7. 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.

8. 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.
9. 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.
10. Хромов А. Б. Пятифакторный опросник личности: учеб.-метод. пособие / А. Б. Хромов. – Курган: Изд-во Курган. гос. ун-та, 2000. – 23 с.
11. Райгородский Д. Я. Практическая психодиагностика. Методики и тесты: учеб. пособие / Д. Я. Райгородский. – Самара: БАХРАХ, 1998. – 672 с.