Психология. Журнал Высшей школы экономики.
2021. Т. 18. № 1. С. 129–144. DOI: 10.17323/1813-8918-2021-1-129-144

Статьи

ИНТЕЛЛЕКТ РЕГИОНОВ РОССИИ СКВОЗЬ ПРИЗМУ СОЦИАЛЬНЫХ СЕТЕЙ

Е.А. ВАЛУЕВА[a], Е.М. ЛАПТЕВА[a], А.А. ГРИГОРЬЕВ[a]

[a] *ФГБУН «Институт психологии Российской академии наук», 129366, Москва, ул. Ярославская, д. 13, к. 1*

Резюме

В работе исследуются связи характеристик текстовых сообщений пользователей социальной сети ВКонтакте с интеллектом. Анализ проводится на региональном уровне: единицами анализа являются регионы Российской Федерации, сопоставляются усредненные показатели сообщений пользователей, проживающих в регионах, с оценками интеллектуального потенциала регионов. Рассмотренные характеристики сообщений включали формальные и грамматические показатели, а также эмоциональные компоненты сообщений. Оценки интеллектуального потенциала регионов были получены путем усреднения региональных результатов тестирования интеллекта мужчин, желающих поступить на военную службу по контракту, и данных об образовательных достижениях регионов (средний результат ЕГЭ лиц, поступивших в российские вузы в 2018 г.). Анализ показал, что четыре формальных и грамматических показателя (средняя длина слова, относительное количество вводных слов и словосочетаний, относительное количество простых непроизводных предлогов и средняя длина предложения), которые могут рассматриваться как маркеры когнитивной сложности текста, вносят независимый вклад в предсказание регионального интеллекта и в совокупности объясняют 60% его дисперсии. Эмоциональные компоненты сообщений не вносят независимого от маркеров когнитивной сложности текста вклада в предсказание регионального интеллекта. Кроме того, была показана связь регионального интеллекта с активностью и грамотностью пользователей социальной сети ВКонтакте. Значение полученных результатов заключается в том, что они ведут к созданию средств оценки интеллекта по тексту, которые будут иметь, можно полагать, определенные преимущества перед традиционными психометрическими методами его измерения в исследованиях региональных различий интеллекта и его изменения во времени.

Ключевые слова: интеллект, цифровые следы, социальная сеть, ВКонтакте, когнитивная сложность, автоматический анализ текста, эмоциональная окраска текста, регионы России.

Работа выполнена при поддержке РФФИ, проект № 17-29-09155 офи_м.

Введение

В последнее время все больший интерес в психологии вызывает исследование активности пользователей в социальных сетях. Привлекательность этого материала заключается в его несомненной экологической валидности: общение и самовыражение в интернет-пространстве происходят естественным образом, поэтому могут считаться идеальным объектом для психологического исследования. На настоящий момент показано, что цифровые следы человека (т.е. совокупность его действий в сети Интернет) могут быть в достаточной степени хорошими предикторами личностных черт Большой пятерки (в особенности — экстраверсии и открытости опыта), а также некоторых эмоциональных состояний — тревоги, депрессии (Латынов, Овсянникова, 2020). При этом наиболее часто анализируется лексика текстовых сообщений пользователей, «лайки», а также визуальный материал — аватарки (фото пользователей) и размещаемые в постах фото или картинки. Обращает на себя внимание тот факт, что в исследованиях цифровых следов человека не очень популярно направление, связанное с изучением когнитивных способностей. Лишь в одной из работ М. Косински с соавт. было показано, что интеллект пользователей Фейсбука может быть предсказан лайками определенных категорий (например, люди с высоким интеллектом лайкают категории «Властелин колец», а с низким — «Харлей Дэвидсон») (Kosinski et al., 2013). Некоторое время назад также были популярны работы, в которых изучались текстовое поведение в социальных сетях, а именно использование текстизмов (например, «4 you», «2 you», «2 B or not to be»), и связь текстового поведения с навыками чтения, грамотностью и т.д. Было показано, что, несмотря на популярность в англоязычных СМИ мнения об увлечении текстизмами как об опасной тенденции для литературного развития и грамотности детей, исследования не подтверждают эту точку зрения (Wood et al., 2014). Результаты показывают, что дети, использующие больше текстизмов, демонстрируют лучшие способности в вербальных рассуждениях и орфографии, лучше читают, имеют больший словарный запас (Plester et al., 2008, 2009). Стоит отметить, однако, что исследования на других возрастных группах (подростки, взрослые) не дают столь однозначной картины (Waldron et al., 2015).

Исследования, имеющие истоки в психолингвистике, демонстрируют, что тексты могут быть охарактеризованы с точки зрения их когнитивной сложности, которая, в свою очередь, может предсказывать уровень интеллектуальных способностей как читателя, для которого этот текст создан (Валуева и др., 2017), так и самого автора (Smirnov, 2017). В работе И. Смирнова в качестве единственного показателя сложности текста была взята средняя длина слова в текстовых сообщениях пользователей ВКонтакте. Средняя длина слова хорошо дифференцировала выпускников высоко- и низкорейтинговых школ Санкт-Петербурга и коррелировала с возрастом пользователя в момент написания поста.

В работе Валуевой с соавт. был выявлен целый ряд показателей, отражающих когнитивную сложность текстов детской художественной литературы.

Среди них были формальные показатели, характеризующие объем текста (длина слов, длина предложений, количество запятых и т.д.), а также морфологические (использование разных типов слов и частей речи) и синтаксические компоненты (типы предложений), позволяющие проанализировать изменения, происходящие с текстом по мере взросления читательской аудитории.

Упомянутые работы И. Смирнова (2017) и Е. Валуевой с соавт. (2017) являются примером принципиально нового подхода к исследованию связи когнитивных способностей и особенностей текстов. Связи показателей в них изучаются не на индивидуальном, а на групповом уровне. Начало интенсивным исследованиям в этом направлении положила первая книга Р. Линна и Т. Ванханена «IQ и благосостояние наций» (Lynn, Vanhanen, 2002), в которой было показано, что интеллект страны (национальный IQ) коррелирует со множеством социально-экономических показателей, характеризующих благосостояние стран. В значительной части работ по поиску социально-экономических коррелятов интеллекта сравниваются показатели стран между собой. Однако существуют и работы на уровне регионов одной страны, демонстрирующие сходные результаты. Подобные исследования проведены по регионам Франции, США, Италии, Португалии, Испании, Китая, Японии, Финляндии, Индии, Турции, России (см.: Grigoriev et al., 2016).

Настоящее исследование также основано на изучении больших групп населения. В нашем случае единицами анализа являются регионы Российской Федерации. Цель работы состояла в том, чтобы изучить взаимосвязь особенностей текстового поведения пользователей социальной сети ВКонтакте с интеллектуальным потенциалом регионов России.

Методика

Сбор данных из социальной сети ВКонтакте

Данные для анализа были загружены из популярной в России социальной сети ВКонтакте, которая предоставляет программный интерфейс (API) для взаимодействия и загрузки ресурсов на программном уровне. Загружались сообщения пользователей, которые указали в профиле возраст от 18 лет и хотя бы одно учебное заведение, а также имеют хотя бы одного друга из указанного учебного заведения. Последнее требование было сформулировано для того, чтобы по возможности отсеять «фейковые» аккаунты (Смирнов и др., 2016). Алгоритм сбора обращался к сайту vk.com с запросом, содержащим название региона и города, и выбирал случайным образом пользователей, имеющих не менее 10 общедоступных публикаций, начиная отсчет от самых свежих публикаций (2019) и до начала 2012 г. Считались только публикации, содержащие текст (т.е. исключались записи с приложениями без комментариев от владельца исследуемого профиля) и не содержащие ссылку на приложение ВКонтакте. Сбор данных по региону останавливался, когда для региона было набрано 3200 пользователей, либо раньше, если исчерпывались доступные профили пользователей, соответствующие условиям.

Помимо самих текстов сообщений (постов), для каждого пользователя были также загружены данные о возрасте, поле, числе друзей, а для каждого сообщения — тип приложения (нет, фото, аудио, видео и т.д.) и количество его просмотров (для записей начиная с 1 января 2017 г.).

Предварительная обработка данных

Первоначально были загружены сообщения 315 441 пользователей старше 18 лет из 85 регионов России, всего 5 457 945 непустых (т.е. содержащих какой-либо текст) сообщений (постов).

Часть сообщений имела признаки автоматического генерирования приложениями ВКонтакте или посторонними сайтами. Например, «Нажми на ссылку, чтобы вступить в мой клан...», «Клевер — ежедневная онлайн-викторина...», «У меня новый Уровень Азарта и куча бонусов...» и т.д. Часть сообщений содержала только гиперссылку, без каких-либо пояснений пользователя. Такие сообщения мы исключили из анализа (всего было установлено около 150 различных признаков, по которым сообщения исключались из анализа). Также были исключены данные пользователей, отметивших в профиле возраст более 70 лет. Анализ данных этих профилей показал, что тексты сообщений демонстрируют случайным образом распределенные результаты всех показателей. Мы предположили, что эти пользователи в большинстве своем являются школьниками, которым ВКонтакте не позволяет указать свой реальный возраст.

В результате в анализ вошли данные 298 806 пользователей и 4 600 023 сообщений этих пользователей. На каждого пользователя приходится от 1 до 333 записей с медианой 8 записей у пользователя. Общее число слов в записях пользователя варьировало от 1 до 146 729 (медиана — 55).

Для дальнейшего анализа данные всех пользователей из региона были объединены в один массив. Все показатели для каждого региона вычислялись на основе этого объединенного массива. Количество постов, вошедших в массив, варьировало в зависимости от региона от 8197 (Чукотский автономный округ) до 150 379 (г. Москва).

Анализируемые показатели

Показатели текстовых сообщений

Анализ данных проводился при помощи кода, написанного на языке R.

В сообщениях пользователей было выделено несколько частей:

1. Словесная часть. Под словом понимался набор кириллических или латинских букв (и дефиса), идущих непрерывно.

2. Эмотиконы (смайлы), т.е. набор типографических знаков, изображающих эмоцию.

3. Хештеги, т.е. набор символов (одно или несколько слов, идущих непрерывно), начинающийся со знака #. Хештеги были удалены из текста и не использовались в дальнейшем анализе.

Хештеги и неэмотиконы из категории Юникода «Other symbol» были удалены из текста и не использовались в дальнейшем анализе.

Словесная часть

В словесной части сообщений была произведена оценка нескольких показателей. За основу были взяты показатели, которые использовались в статье Е.А. Валуевой с соавт. (Валуева и др., 2017), однако специфика текстов в социальных сетях не позволила абсолютно точно воспроизвести все переменные. Анализировались следующие характеристики.

I Формальные характеристики, связанные с общим объемом текстовой продукции:

1) число слов в одной записи (медиана по региону от среднего числа слов по записям пользователя), далее — средняя длина поста;

2) длина слова (медиана по региону от средней длины слова у пользователя);

3) количество точек в тексте (по отношению к общему количеству слов). Мы полагаем, что данный показатель, хотя и весьма приблизительно в случае с текстами в социальных сетях, отражает среднюю длину предложения (чем больше точек на слово, тем меньше длина предложения). Приблизительность оценки длины предложения связана с тем, что, как правило, пользователи не ставят точки (и не используют заглавных букв), если пишут одно короткое предложение, или могут применять эмотиконы и/или хештеги вместо знаков препинания;

4) количество запятых в тексте (по отношению к общему количеству слов).

II Морфологические компоненты:

5) количество вводных слов и словосочетаний (относительно общего количества слов в тексте). Примерами вводных слов являются такие, как «вероятно», «конечно», «видимо» и т.д. Предполагается, что вводные слова имеют множество особенностей и функций, усложняющих структуру высказывания;

6) количество компаративов, т.е. прилагательных и наречий в сравнительной степени (относительно общего количества слов в тексте), например: «дальше», «ближе», «веселее», «проще» и т.д. Слова, выражающие степень сравнения, вносят дополнительное логическое звено в характеристику объектов действительности, требуют осознания непростых взаимоотношений между ними, поэтому могут служить маркерами усложнения когнитивной организации;

7) количество простых непроизводных предлогов (относительно общего количества слов в тексте), например: «в», «за», «на», «под», «перед» и т.д. Грамматически предлоги отвечают за бóльшую детализацию и визуализированность текста по сравнению с элементарными субъектно-предикатными

предложениями типа «Мама мыла раму», поэтому их наличие может рассматриваться как маркер усложнения текста;

8) количество модальных частиц (относительно общего количества слов в тексте), например: «же», «неужели», «разве» и т.д. Модальные частицы вносят в предложение эмоционально-оценочные, вопросительные, уступительные и другие семантические оттенки, поэтому могут являться маркерами когнитивного усложнения текста.

III Синтаксические компоненты:

9) количество вопросительных и 10) количество восклицательных предложений (относительно общего количества слов в тексте). Количество восклицательных предложений оценивалось по количеству восклицательных знаков. Количество вопросительных предложений рассчитывалось как среднее z-значений двух показателей (коррелирующих на уровне 0.36) — количество знаков вопроса и количество вопросительных слов («какой», «куда», «откуда», «насколько» и т.д.).

IV Эмоциональная окраска слов

Для оценки эмоциональной окраски слов была использована база данных ENRuN (Люсин, Сысоева, 2017), содержащая 378 существительных русского языка, оцененных по эмоциональным категориям «радость», «грусть», «злость», «страх» и «отвращение». Четыре слова были исключены из анализа, поскольку их словоформы совпадали с другими словами (например, «горе» может означать название чувства, а может быть падежной формой слова «гора»). Был произведен поиск слов из словаря ENRuN (с учетом всех словоформ) в сообщениях пользователей ВКонтакте, и на основе этих совпадений был вычислен средний балл эмоциональной окраски слов по пяти категориям. Так как оценки по категориям высоко коррелировали между собой (факторный анализ продемонстрировал наличие одного фактора, объясняющего более 90% дисперсии), в качестве общего показателя были взяты факторные оценки по первому фактору. Оценки были инвертированы (умножены на −1) так, чтобы высокие значения соответствовали позитивной окрашенности слов, а низкие — негативной.

Эмотиконы

На сайте https://unicode-table.com/ были выбраны эмотиконы из основного набора. Выбранные эмотиконы мы разделили на три категории:

1) позитивные (N = 25) — эмотиконы с различными вариантами улыбок;

2) негативные (N = 30) — эмотиконы с прямыми или опущенными уголками «губ», оскалом или круглым ртом, с графическими признаками страха;

3) жестовые (N = 16) — изображения действий — губы, сложенные в поцелуй, руки в положении мольбы, закрытый руками рот, аплодисменты и т.д.

Показатели интеллектуального потенциала регионов РФ

Для подсчета показателя интеллектуального потенциала регионов России (регионального интеллекта) были использованы две переменные. Во-первых, из работы К. Сугоняева с соавт. (Sugonyaev et al., 2018) были взяты оценки интеллекта регионов, полученные путем тестирования интеллекта мужчин, желающих поступить на военную службу по контракту (тестирование проводилось на сайте Министерства обороны РФ, выборка составила 238 619 человек). Во-вторых, на основе данных, представленных на сайте Высшей школы экономики (https://ege.hse.ru/stata_2018), была произведена оценка интеллекта регионов по среднему баллу ЕГЭ лиц, поступивших в российские вузы в 2018 г. Оценка производилась следующим образом. В данных, представленных на сайте Высшей школы экономики, сообщены средние баллы ЕГЭ поступивших в каждый вуз на обучение по каждой специальности. Первым шагом было усреднение внутри регионов оценок лиц, поступивших на обучение по данной специальности, по вузам, со взвешиванием на численность поступивших. Таким образом, были получены средние оценки лиц, поступивших в данном регионе на обучение по данной специальности. Специальности различались по престижности. Например, в Алтайском крае средний балл ЕГЭ поступивших на обучение по специальности «Государственное и муниципальное управление» был 76.2, а поступивших на обучение по специальности «Вооружение» — всего 52.3; в Волгоградской области средний балл поступивших на эти две специальности составил 73.6 и 57.1 соответственно. Между тем вузы, обучающие по тем или иным специальностям, распределены по регионам неравномерно. Чтобы исключить этот источник вариации, на втором шаге баллы ЕГЭ были стандартизированы внутри специальностей. На третьем шаге стандартизированные оценки были усреднены по специальностям, в результате были получены региональные оценки интеллекта.

Так как два показателя достаточно высоко коррелировали между собой ($r = 0.62$), в качестве значения регионального интеллекта было взято среднее z-оценок по этим показателям. Лишь для одного из 85 регионов РФ (Ненецкий АО) не была представлена ни одна из оценок (в регионе отсутствуют вузы, а прошедших тестирование на сайте Министерства обороны оказалось слишком мало (менее 100 человек), чтобы надежно оценить уровень интеллекта в регионе). В связи с этим весь анализ взаимосвязей регионального интеллекта с особенностями сообщений в ВКонтакте проводился для 84 регионов РФ.

Результаты

В таблице 1 приведены корреляции регионального интеллекта с выделенными показателями текстовых сообщений ВКонтакте.

Перечисленные показатели можно объединить в две большие группы — показатели когнитивной сложности сообщений и показатели эмоциональной

Таблица 1

Корреляции регионального интеллекта с показателями текстовых сообщений ВКонтакте (коэффициенты корреляции Пирсона)

Показатель	*r*	*p*
Среднее количество слов в посте	0.39	0.000
Средняя длина слова	0.64	0.000
Количество точек (длина предложений)	−0.23	0.034
Количество запятых	0.24	0.032
Вводные слова	0.44	0.000
Компаративы	0.44	0.000
Непроизводные предлоги	0.62	0.000
Частицы	0.09	0.434
Вопросительные предложения	−0.23	0.033
Восклицательные предложения	0.39	0.000
Эмоциональная окраска слов	0.58	0.000
Позитивные эмотиконы	0.06	0.599
Негативные эмотиконы	0.25	0.021
Эмотиконы-жесты	0.52	0.000

составляющей сообщений пользователей ВКонтакте. Результаты будут представлены по каждой группе показателей отдельно.

Связь регионального интеллекта с особенностями когнитивной составляющей сообщений пользователей ВКонтакте

Из таблицы 1 видно, что показатели регионального интеллекта демонстрируют значимые и достаточно высокие корреляции с формальными и морфологическими характеристиками сообщений пользователей ВКонтакте. Наиболее высокая корреляция наблюдается со средней длиной слова (0.64), что неудивительно, так как длина слова является традиционной мерой сложности текста (Smirnov, 2017; Yasseri et al., 2012). Также положительные корреляции с региональным интеллектом наблюдаются для среднего количества слов в посте (0.39), количества запятых (0.24) и точек (−0.24)[1], вводных слов (0.44), компаративов (0.44) и непроизводных предлогов (0.62). Единственным показателем из словесной части сообщений, который, вопреки нашим ожиданиям, оказался не связанным с региональным интеллектом, явилось количество употребляемых частиц (0.09). Прямой пошаговый регрессионный анализ показал, что четыре показателя (длина слов, вводные слова, непроизводные предлоги и длина предложений) в совокупности объясняют 60% дисперсии регионального интеллекта (таблица 2). Остальные предикторы (количество слов в посте, количество запятых, компаративы) не вносили значимого вклада в регрессионную модель.

[1] Отрицательная корреляция свидетельствует о том, что более высокий интеллект связан с порождением более длинных предложений.

Таблица 2

Результаты регрессионного анализа (зависимая переменная — региональный интеллект, независимые переменные — формальные и морфологические особенности сообщений пользователей)

Предикторы	B	Std. Error	β	t	p	R^2 change*
(Constant)	−40.73	5.86		−6.96	0.000	
Длина слов	6.88	1.26	0.46	5.46	0.000	0.41
Вводные слова	711.48	185.35	0.28	3.84	0.000	0.13
Непроизводные предлоги	72.62	24.02	0.26	3.02	0.003	0.05
Количество точек (длина предложений)	−65.24	25.76	−0.18	−2.53	0.013	0.03

* Изменение R^2 указано для последовательных моделей, включающих в себя соответствующую переменную.

Из таблицы 1 также следует, что количество вопросительных предложений отрицательно коррелирует с интеллектом, а количество восклицательных предложений — положительно. Это противоречит результатам, полученным в исследовании Е.А. Валуевой с соавт. (Валуева и др., 2017), где было показано, что по мере роста когнитивных способностей предполагаемой читательской аудитории текста увеличивается количество вопросительных предложений и уменьшается количество восклицательных. Так как процитированные данные получены на принципиально другом материале (тексты детской художественной литературы), можно предположить, что вопросительные и восклицательные знаки в сообщениях социальных сетей не являются характеристикой когнитивной сложности текста, а отражают эмоциональную составляющую сообщения. Это предположение будет проверено в следующем разделе.

Эмоциональная составляющая сообщений пользователей ВКонтакте и ее связь с региональным интеллектом

В таблице 1 показано, что эмоциональная окраска используемых слов коррелирует с региональным интеллектом на достаточно высоком уровне (0.58). Также с региональным интеллектом коррелируют, как уже было сказано, использование вопросительных (−0.23) и восклицательных (0.39) знаков, негативных (0.25) и жестовых (0.52) эмотиконов. Использование позитивных эмотиконов (улыбающихся рожиц) никак не связано с интеллектом. В таблице 3 представлена корреляционная матрица «эмоциональных» переменных. Можно заметить, что паттерн корреляций данного набора переменных с эмоциональной окраской слов сходен с паттерном корреляций этих переменных с региональным интеллектом (отрицательная корреляция вопросительных предложений, отсутствие корреляции с позитивными эмотиконами, положительные корреляции — со всеми остальными).

Таблица 3

Корреляционная матрица эмоциональных компонентов

	Эмоциональная окраска слов	Вопросительные предложения	Восклицательные предложения	Позитивные эмотиконы	Негативные эмотиконы
Вопросительные предложения	−0.429**	-			
Восклицательные предложения	0.728**	−0.213*	-		
Позитивные эмотиконы	0.151	0.052	−0.060	-	
Негативные эмотиконы	0.424**	−0.181	0.033	0.354**	-
Эмотиконы-жесты	0.639**	−0.359**	0.356**	0.200	0.681**

* $p < 0.05$, ** $p < 0.01$.

Таблица 4

Результаты регрессионного анализа (зависимая переменная — региональный интеллект, независимые переменные — эмоциональные компоненты сообщений пользователей)

Предикторы	B	Std. Error	β	t	p	R^2 change*
(Constant)	−0.547	0.261		−2.100	0.039	
Эмоциональная окраска слов	0.383	0.105	0.418	3.643	0.000	0.333
Эмотиконы-жесты	5.176	2.381	0.249	2.174	0.033	0.037

* Изменение R^2 указано для последовательных моделей, включающих в себя соответствующую переменную.

Прямой пошаговый регрессионный анализ показал, что две переменные (эмоциональная окраска слов и эмотиконы-жесты) объясняют 36% дисперсии регионального интеллекта (см. таблицу 4), а все остальные переменные (вопросительные и восклицательные предложения, негативные эмотиконы) не вносят в модель значимого вклада.

По всей видимости, интеллект связан с эмоциональностью текстов и в силу этого — с использованием восклицательных и вопросительных предложений (знаков) как с дополнительными средствами выражения эмоций, помимо окраски слов. При этом примечательно, что восклицательные предложения передают позитивные эмоциии, а вопросительные — с негативные.

Затем мы проверили, вносят ли эмоциональные компоненты сообщений пользователей дополнительный вклад (по сравнению с когнитивными компонентами) в объяснение региональных различий в интеллекте. Результаты регрессионного анализа показали, что при контроле когнитивных компонентов вклад эмоциональных не значим (p(F change) = 0.515) (таблица 5).

Таблица 5

Результаты регрессионного анализа (зависимая переменная — региональный интеллект, независимые переменные — когнитивные и эмоциональные компоненты сообщений пользователей)

Предикторы	B	Std. Error	β	t	p
(Constant)	−36.93	6.76		−5.46	0.000
Длина слов	6.24	1.38	0.41	4.51	0.000
Вводные слова	689.02	187.51	0.27	3.67	0.000
Непроизводные предлоги	60.68	26.31	0.22	2.31	0.024
Количество точек (длина предложений)	−5.38	27.78	−0.15	−1.99	0.050
Эмоциональная окраска слов	0.08	0.10	0.08	0.80	0.427
Эмотиконы-жесты	0.91	2.02	0.04	0.45	0.653

Связь срегионального интеллекта с общей активностью и грамотностью пользователей ВКонтакте

Алгоритм загрузки данных позволил нам оценить параметр, который мы назвали общей активностью пользователей. Как было указано выше, для каждого пользователя загружались сообщения, начиная с самых новых и заканчивая хронологически более ранними. На основании даты публикации каждого сообщения нами был посчитан средний год публикации. Этот показатель (средний год публикации сообщений в периоде с 2012 по 2019 г.) свидетельствует об общей активности пользователей в социальной сети, так как наиболее активные пользователи будут иметь большее количество более свежих записей по сравнению с менее активными. Корреляция регионального интеллекта со средним годом публикации составила 0.77 ($p < 0.0001$). Частично эту связь можно объяснить увеличением длины слов в сообщениях, происходящим с течением времени (Smirnov, 2017), однако даже парциальная корреляция интеллекта с активностью пользователей при контроле длины слов осталась достаточно высокой и значимой ($r = 0.58$, $p < 0.0001$).

Еще один параметр, который мы косвенно смогли оценить на основе полученных данных, — это грамотность пользователей. При оценке эмоциональной окраски сообщений мы производили подсчет частоты встречаемости слов из словаря ENRuN в сообщениях пользователей. Количество совпадений слов в сообщении со словами из словаря может рассматриваться как косвенная мера грамотности пользователей. Если слово не совпадает со словарем, это может быть связано как с тем, что его нет в тексте в принципе, так и с тем, что оно написано неправильно. В силу того что у нас нет особых оснований полагать, что распределение встречаемости слов, фиксируемых словарем ENRuN, не является равномерным в регионах, мы считаем, что разброс по этой переменной обусловлен правильным или неправильным написанием слов (т.е. грамотностью

пользователей). Корреляция общего количества совпадений с эмоциональным словарем (при контроле общего количества слов в регионе) составила 0.41 ($p < 0.0001$).

Обсуждение

Первый момент, на который хотелось бы обратить внимание в обсуждении, — это вопрос о пригодности данных из социальных сетей для автоматического анализа.

Алгоритм сбора данных на этапе загрузки сообщений отфильтровывал пользователей по следующим параметрам: 1) пользователи, зарегистрированные в одном из 85 регионов РФ; 2) пользователи, имеющие общедоступный профиль; 3) пользователи старше 18 лет; 4) пользователи, у которых указано хотя бы одно учебное заведение (школа или вуз); 5) пользователи, имеющие хотя бы одного друга из указанного учебного заведения; 6) пользователи, имеющие хотя бы одно непустое текстовое сообщение на своей странице. Чтобы собрать нашу выборку (315 441 пользователей), пришлось отсеять около 85 млн пользователей, которые не соответствовали перечисленным критериям отбора. Это составляет примерно 1/7 всех пользователей ВКонтакте (на момент написания статьи общее количество зарегистрированных пользователей ВКонтакте составляло примерно 592.4 млн). На этапе обработки данных мы отсеяли еще около 1/5 собранных постов, которые оказались не пригодны для нашего анализа (автоматически сгенерированные ссылки и тексты, посты, состоящие из 1 буквы, хештеги и т.д.).

Однако, несмотря на использование достаточно строгих критериев отбора, удалось собрать выборку, в которой были хорошо представлены все регионы страны. Таким образом, получение валидных данных из социальных сетей возможно, но требует организации правильной процедуры автоматического отбора.

Во-вторых, центральным в данном исследовании является вопрос о пригодности данных из социальных сетей для оценки интеллекта населения регионов РФ. Значимые корреляции между рядом показателей текстовых сообщений и оценками интеллекта дают основания для положительного ответа на этот вопрос. Это укрепляет оптимизм, который вселяют результаты недавних исследований (Валуева и др., 2017; Woodley of Menie et al., 2015), относительно перспектив создания средств оценки интеллекта, в частности группового, по текстам. Важность появления таких средств трудно переоценить. Во-первых, не секрет, что оценивание регионального интеллекта, его динамики с помощью психометрических тестов зачастую связано с большими организационными трудностями. Психологам нередко приходится полагаться на результаты исследований образовательных достижений, которые не везде бывают легко доступны и могут быть далеко не столь валидной мерой регионального интеллекта, как это принять думать. Появление средств диагностики интеллекта по легко доступному материалу — текстам, нужные параметры выборок которого исследователи смогут обеспечить, позволит резко

увеличить количество и улучшить качество оценок регионального интеллекта. Во-вторых, как отмечают М. Вудли оф Мени с соавт. (Woodley of Menie et al., 2015), частота правильных ответов на пункты психометрических тестов с вариантами ответа (а таких большинство) увеличивается за счет использования респондентами тактик угадывания. Из текстов же можно извлечь более экологически валидные данные о изменении интеллекта во времени (Ibid.), что обуславливает значение разработки средств оценки интеллекта по текстам для исследований эффекта Флинна.

Полученные результаты довольно близко подводят к созданию одного из средств оценки интеллекта по тексту — опирающегося на формальные и грамматические показатели текста. Четыре таких показателя (средняя длина слова, количество вводных слов и словосочетаний относительно общего количества слов в тексте, количество простых непроизводных предлогов относительно общего количества слов в тексте и средняя длина предложения, оцениваемая по количеству точек по отношению к общему количеству слов) в совокупности объяснили 60% дисперсии регионального интеллекта. Эти показатели могут рассматриваться как маркеры когнитивной сложности текста, что и обусловило их использование в данном исследовании. Дальнейшая работа в этом направлении будет заключаться, во-первых, в оптимизации состава таких показателей, а во-вторых, в сопоставлении оценок регионального интеллекта с социально-экономическими показателями регионов РФ.

Проведенное исследование показало также неэффективность эмоциональных компонентов сообщений как независимых от маркеров когнитивной сложности текста предикторов регионального интеллекта. Этот негативный результат, однако, ставит вопрос об источниках корреляций этих компонентов с региональным интеллектом. Кроме того, показана связь активности и грамотности пользователей социальной сетью ВКонтакте с региональным интеллектом. Связь с активностью представляет собой, вероятно, одну из трудно-объяснимых связей интеллекта (другим примером является его связь с политическими ориентациями), что же касается связи с грамотностью, то этот результат, не являясь, напротив, неожиданным (см., например: Григорьев и др., 2015), нуждается в дальнейшей проверке.

Выводы

По результатам проведенного исследования можно сделать следующие выводы.

1. Социальные сети являются источником пригодных для автоматического анализа данных.

2. Формальные и грамматические показатели текста целесообразно использовать при создании средств оценки регионального интеллекта по текстам.

3. Эмоциональные компоненты сообщений нецелесообразно использовать для оценки регионального интеллекта по текстам.

Литература

Валуева, Е. А., Данилевская, Н. М., Лаптева, Е. М., Ушаков, Д. В. (2017). Когнитивная сложность художественных текстов для детей: квантитативные методы оценки. *Вопросы психолингвистики, 31*(1), 42–61.

Григорьев, А. А., Лаптева, Е. М., Ушаков, Д. В. (2015). Образовательные достижения районов Московской области воспроизводят уровень грамотности в XIX в. *Сибирский психологический журнал, 56*, 69–85.

Латынов, В. В., Овсянникова, В. В. (2020). Прогнозирование психологических характеристик человека на основании его цифровых следов. *Психология. Журнал Высшей школы экономики, 17*(1), 166–180.

Люсин, Д. В., Сысоева, Т. А. (2017). Эмоциональная окраска имен существительных: база данных ENRuN. *Психологический журнал, 38*(2), 122–131.

Смирнов, И. Б., Сивак, Е. В., Козьмина, Я. Я. (2016). В поисках утраченных профилей: достоверность данных «ВКонтакте» и их значение для исследований образования. *Вопросы образования*, *4,* 106–122.

Ссылки на зарубежные источники см. в разделе References после англоязычного блока.

Валуева Екатерина Александровна — научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии Российской академии наук», кандидат психологических наук.
Сфера научных интересов: когнитивная психология, интеллект, творчество.
Контакты: ekval@list.ru

Лаптева Екатерина Михайловна — научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии Российской академии наук», кандидат психологических наук.
Сфера научных интересов: кристаллизованный интеллект, потребность в познании, когнитивные и мотивационные механизмы кристаллизованного интеллекта.
Контакты: ek.lapteva@gmail.com

Григорьев Андрей Александрович — ведущий научный сотрудник, лаборатория психологии и психофизиологии творчества, ФГБУН «Институт психологии РАН», доктор филологических наук, доцент.
Сфера научных интересов: интеллект, индивидуальные различия, психолингвистика.
Контакты: andrey4002775@yandex.ru

Regional IQ Differences in Russia through the Prism of Social Media

E.A. Valueva[a], E.M. Lapteva[a], A.A. Grigoriev[a]

[a] *Institute of Psychology, Russian Academy of Sciences, 13 build. 1, Yaroslavskaya Str., Moscow, 129366, Russian Federation*

Abstract

The article examines the relationship between characteristics of text messages, composed by users of social network (VKontakte), and intelligence. The analysis is conducted on the regional level: we compared the regional IQ with the text parameters averaged over the users living in one region. The text parameters include formal, grammatical and emotional indexes. The regional IQ is computed as an average z-score of the Unified State Examination score (high school entrants, 2018) and IQ score of the attendees to the volunteer military service. Four text parameters that can be considered as markers of the text cognitive complexity (mean word length, mean sentence length, percent of the parenthetic words and phrases, percent of the simple propositions) predicted regional IQ independently and explained 60% of its variance. Emotional index correlates with regional IQ, but does not predict regional IQ independently of cognitive complexity markers. Moreover, we revealed correlations between regional IQ and literacy of VKontakte users. The significancy of these results is creating the new IQ measure, which allows evaluating regional IQ and its dynamics by means of text analysis. The method has an advantage over the traditional psychometric IQ measures in this field of research.

Keywords: intelligence, digital footprints, social media, VKontakte, cognitive complexity, text automatic analysis, text sentiment analysis, regions of Russia.

References

Grigoriev, A. A., Lapteva, E. M., & Ushakov, D. V. (2015). Educational performance of Moscow region districts reproduce their literacy level in the XIX century: mechanisms of the "cultural genetics". *Siberian Journal of Psychology, 56*, 69–85. (in Russian)

Grigoriev, A., Ushakov, D., Valueva, E. A., Zirenko, M., & Lynn, R. (2016). Differences in educational attainment, socio-economic variables and geographical location across 79 provinces of the Russian Federation. *Intelligence, 58*, 14–17.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805.

Latyunov, V. V., & Ovsyannikova, V. V. (2020). Predicting psychological characteristics from digital footprints. *Psychology. Journal of Higher School of Economics, 17*(1), 166–180. (in Russian)

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

Lyusin, D. V., & Sysoeva, T. A. Emotional norms for nouns: The database ENRuN. *Psikhologicheskii Zhurnal, 38*(2), 122–131. (in Russian)

Plester, B., Wood, C., & Bell, V. (2008). Txt msg n school literacy: does texting and knowledge of text abbreviations adversely affect children's literacy attainment? *Literacy, 42*(3), 137–144.

Plester, B., Wood, C., & Joshi, P. (2009). Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *The British Journal of Developmental Psychology, 27*(1), 145–161.

Smirnov, I. (2017). The digital Flynn effect: Complexity of posts on social media increases over time. *Lecture Notes in Computer Science, 10540*, 24–30. doi:10.1007/978-3-319-67256-4_3

Smirnov, I., Sivak, E., & Kozmina, Y. (2016). In search of lost profiles: The reliability of VKontakte data and its importance in educational research. *Educational Studies Moscow, 4*, 106–122. (in Russian)

Sugonyaev, K., Grigoriev, A., & Lynn, R. (2018). A new study of differences in intelligence in the provinces and regions of the Russian Federation and their demographic and geographical correlates. *Mankind Quarterly, 59*(1), 31–37.

Valueva, E. A., Danilevskaya, N. M., Lapteva, E. M., & Ushakov, D. V. (2017). Cognitive complexity of children fiction: quantitative methods of evaluation. *Journal of Psycholinguistics, 31*(1), 42–61. (in Russian)

Waldron, S., Kemp, N., Plester, B., & Wood, C. (2015). Texting behavior and language skills in children and adults. In L. D. Rosen, N. A. Cheever, & L. M. Carrier (Eds.), *The Wiley handbook of psychology, technology, and society* (pp. 232–240). John Wiley & Sons, Ltd.

Wood, C., Kemp, N., Waldron, S., & Hart, L. (2014). Grammatical understanding, literacy and text messaging in school children and undergraduate students: A concurrent analysis. *Computers and Education, 70*, 281–290.

Woodley of Menie, M. A., Fernandes, H. B. F., Figueredo, A. J., & Meisenberg, G. (2015). By their words ye shall know them: Evidence of genetic selection against general intelligence and concurrent environmental enrichment in vocabulary usage since the mid 19th century. *Frontiers in Psychology, 6*, 361. doi:10.3389/fpsyg.2015.00361

Yasseri, T., Kornai, A., & Kertész, J. (2012). A practical approach to language complexity: A Wikipedia case study. *PLoS ONE, 7*(11), e48386.

Ekaterina A. Valueva — Research Fellow, Institute of Psychology, Russian Academy of Sciences, PhD in Psychology.
Research Area: cognitive psychology, intelligence, creativity.
E-mail: ekval@list.ru

Ekaterina M. Lapteva — Research Fellow, Institute of Psychology, Russian Academy of Sciences, PhD in Psychology.
Research Area: crystalized intelligence, need for cognition, cognitive and motivational mechanisms of crystallized intelligence.
E-mail: ek.lapteva@gmail.com

Andrei A. Grigoriev — Leading Research Fellow, Institute of Psychology of Russian Academy of Sciences, DSc in Philology, Associate Professor.
Research Area: intelligence, individual differences, psycholinguistics.
E-mail: andrey4002775@yandex.ru