МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 2** **Москва 2021**

ОБЩИЙ РАЗДЕЛ

УДК [004.75.056:004.455.1]:316.7

Н.П. Крылова, Е.Н. Левашов

Перспективы блокчейн технологий в информационном обществе

Обоснована актуальность и представлен анализ отечественных и зарубежных исследований сущности блокчейн технологий в современном информационном обществе. Рассматривается появление данной технологии, исследуется механизм ее действия в информационном контексте. Отдельно описана роль информации в блокчейн процессах. Выявлены свойства блокчейн технологий, структурирован мировой опыт в ее применении. Приводятся статистические данные относительно информационной безопасности общества. Дается авторская трактовка перспектив развития технологии блокчейн в информационном обществе.

Ключевые слова: *блокчейн, информация, информационная безопасность, информационное общество, технология, процесс, перспективы*

DOI: 10.36535/0548-0027-2021-02-1

ВВЕДЕНИЕ

Общие тенденции глобализации общества, трансформация информационных процессов, осуществляемых в мире, способствуют появлению революционных технологий, которые затрагивают все цифровое бизнес-пространство. Развитие и стремительное распространение Интернета привели к появлению новых технологий, основанных на блокчейн.

Феномен технологий блокчейн вызывает живой интерес мировой общественности, представителей

бизнеса и науки на разных уровнях – от личного до государственного. Мировое сообщество изучает блокчейн с точки зрения его как общественного феномена, инновационной технологии, как нового способа ведения бизнес-процессов. Несмотря на относительно молодой возраст блокчейн технологий, они привлекают внимание многих исследователей, которые анализируют ее финансовые, правовые аспекты. При этом роль, свойства информации в реализации технологии блокчейн недостаточно изучены, а перспективы этой технологии вызывают неоднозначные мнения в научной среде.

На страницах научных журналов, на полях симпозиумов и конференций учеными, специалистами высказываются противоречивые точки зрения относительно деструктивного и позитивного влияния блокчейн, криптовалюты на мировое развитие экономики, общества. Многими отмечается, что традиционные классические устои общества, формы экономической деятельности претерпели большие изменения с появлением технологий блокчейн.

В мире не существует единого сформировавшегося понимания технологии блокчейн, однако отмечается высокий риск ее реализации. В научно-исследовательской среде появились новые понятия: «блокчейн экономика», «криптоэкономика» – они находятся в стадии своего формирования, но уже противопоставляются классической традиционной экономике. В блокчейн экономике информация приобретает новые материальные свойства, качества, которые пока не исследованы в полной мере.

В контексте блокчейн каждое государство выбирает индивидуальный путь развития, этот процесс еще не закончен, он находится в стадии формирования и активного обсуждения. Различные страны постепенно осваивают данные технологии применительно к своим государственным и правовым стандартам.

Мировое сообщество отмечает, что блокчейн как распределенная база данных обеспечивает новые возможности для хранения и передачи информации. Изучение роли информации в реализации всего процесса блокчейн вызывает особый интерес. С одной стороны, технологии блокчейн рассматриваются как перспектива экономического роста, с другой – ученые отмечают высокие риски внедрения данной технологии [1].

ИСТОРИЯ ВОПРОСА ВОЗНИКНОВЕНИЯ БЛОКЧЕЙН

В соответствии с программой «Цифровая экономика Российской Федерации» одним из стратегических направлений развития экономики в РФ является внедрение цифровых платформ работы с данными и информацией для обеспечения потребностей бизнеса и граждан[1].

Необходимость развития цифровой экономики была озвучена Президентом РФ на заседании Совета по стратегическому развитию и приоритетным проектам. В соответствии с данной тенденцией [2] издан Указ Президента РФ от 9 мая 2017 г. № 203 «О стратегии развития информационного общества в Российской Федерации на 2017-2030 годы»[2].

Рассмотрим этапы поступательного развития технологии блокчейн в мире. Примитивная форма блокчейна известна как хэш-дерево или дерево Меркла (1979 г.). Действие данного механизма было основано на обмене информацией и проверке данных между компьютерными системами.

В 1991 г. впервые была предложена идея цепочек блоков, связанных криптографическими алгоритмами, в форме создания сертификатов, защищающих от подделки временные метки в электронных документах. Спустя год эта идея была дополнена применением деревьев Меркле, что способствовало возможности объединять несколько сертификатов документов в один блок.

В 2008 г. цепочка блоков в виде блокчейна была предложена Сатоши Накамото. В 2009 г. был внедрен первый публичный блокчейн, а выпущенный с его помощью биткоин стал самым популярным нативным активом с использованием криптографии и концепции блокчейн [3].

Технологии блокчейн используются как в функционировании криптовалюты, так и находят свое применение в других сферах. Эта технология основана на том, что имеется определенное количество блоков информации, образующих неразрывную цепочку. Каждый блок включает массив данных, подтверждающих подлинность предыдущего блока, посредством чего цепочку невозможно фальсифицировать. Если заменить один из блоков в цепочке, такая замена сразу будет выявлена [4].

По одной из трактовок сущности блокчейн блок состоит из «тела» и «заголовка», из них составляется последовательная цепочка. Каждый блок имеет свой уникальный ключ, посредством которого обеспечивается взаимосвязь блоков, так как в заголовке каждого блока содержится ключ от предыдущего. Информацию, хранящуюся в блоках цепочки, могут получать пользователи сети, имеющие доступ к ней. Доступ открывает специальный закрытый ключ на основе криптографического алгоритма. Это обеспечивает высокий уровень безопасности проводимых операций (транзакций), невмешательство в операции, невозможность изменить или отменить уже совершенные операции (транзакции) [5, 6].

В табл. 1 представлены различные трактовки понятия блокчейн.

Сущность технологии блокчейн представляется в виде последовательности действий:

- во все узлы сети отправляется транзакция (операция);
- узлы сети добавляют данные в определенный блок;
- узлы проверяют выполнение определенного условия, заданного разработчиками;

[1] Распоряжение Правительства РФ от 28.07.2017 № 1632-р Об утверждении программы «Цифровая экономика Российской Федерации». – URL: http://base.garant.ru/71734878/ (дата обращения: 10.12.2020)

[2] Указ Президента РФ от 9 мая 2017 г. № 203 «О стратегии развития информационного общества в Российской Федерации на 2017-2030 годы». – URL: http://www.garant.ru/products/ipo/prime/doc/71570570/ (дата обращения: 10.12.2020)

- если условие выполняется, то блок данных отправляется всем участникам сети;
- блок данных проходит дальнейшую проверку;
- в случае прохождения проверки блок добавляется в цепочку [15].

Выделяются две группы пользователей блокчейн:

1) стандартные – создают записи, проводят транзакции (операции);

2) майнеры – собирают информацию, проверяют ее, объединяют в блоки и распространяют по сети пользователям [5].

В то же время отмечаются три группы участников блокчейн:

- пользователи – владельцы электронных кошельков, хранящие криптовалюту и осуществляющие переводы другим пользователям;
- майнеры – участники, осуществляющие обработку транзакций пользователей и подбирающие ключ (хеш) для формирования блоков;
- серверы – участники, осуществляющие хранение общей цепочки блокчейн и операции по проверке правильной последовательности блоков [16].

Таблица 1

Трактовки понятия блокчейн и роль информации в ней

Автор (источник)	Определение
Н.Г. Леонова [7]	Выстроенная последовательная цепочка блоков, содержащих определенную информацию
	Распределенная база данных, которая имеет множество копий
Г.О. Крылов, А.В. Токолов [8]	Распределенная база данных, содержащая данные обо всех транзакциях, проведенных участниками системы. Информация хранится в виде цепочки блоков, в которых записано определенное число транзакций. Транзакция в блокчейне – информация, проверяемая участниками системы и встраиваемая в цепочку
С.А. Попов, И.С. Охрицкий [4]	Построенная по определенным правилам непрерывная последовательная цепочка блоков, содержащих информацию, и связанная криптографическими алгоритмами. Каждый блок включает идентификатор, криптографически закодированную ссылку на предыдущий блок и набор данных
И.В. Юшин [5]	Децентрализованная база данных, основанная на одноранговой сети, общем реестре и криптографии публичного и приватного ключа
А.А. Рязанова [9]	Непрерывная цепочка информационных блоков, выстроенных в определенной последовательности, обладающая свойствами неразрывности и прочности. Неразрывность означает, что блоки выстраиваются в строго определенной последовательности. Прочность – невозможность удаления, замены, фальсификации блока из цепочки
Е.В. Сазанова [10]	Непрерывная последовательная цепочка блоков, построенная по определенным правилам и распределенно хранящаяся на множестве компьютеров, каждый блок содержит определенную информацию, временную метку и ссылку на предыдущий блок
	Распределенная база данных, хранящая постоянно растущий список упорядоченных записей
Г.Г. Амиргамзаев, Х.А. Магомедова [11]	Неизменяемая структура данных, включающая списки блоков, в которой каждый следующий блок содержит хэш предыдущего блока. В результате хэширования цепочка блоков становится неизменяемой, нельзя изменить или удалить блок из цепочки, не изменив остальные блоки
Н.Ю. Романенко, О.В. Степнова [12]	Полностью распределенная пиринговая система журналов учета, использующая алгоритм обработки информации последовательно расположенных и взаимосвязанных между собой блоков данных как единого целого посредством криптографических технологий защиты данных для обеспечения целостности системы
О.В. Харченко [13]	Многофункциональная и многоуровневая информационная технология, предназначенная для надежного учета различных активов, может быть средством регистрации, учета и обмена финансовых, материальных и нематериальных активов
Криптовалюты и блокчейн как атрибуты новой экономики [14]	Построенная на основе заданных алгоритмов в распределенной базе данных последовательность взаимосвязанных блоков с информацией о проведенных операциях (транзакциях)

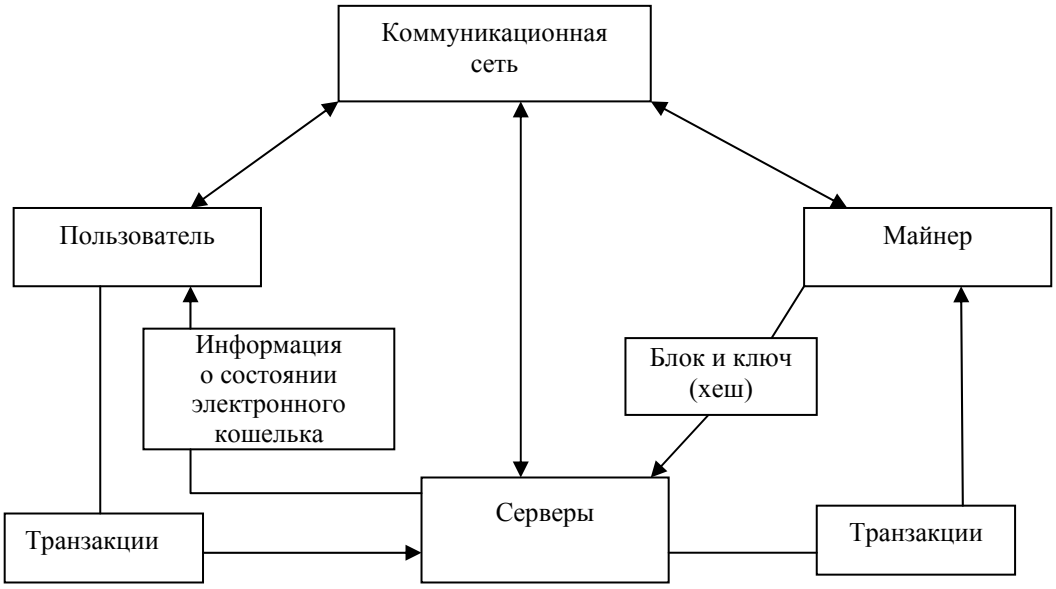

Рис. 1. Взаимодействие участников при обработке транзакций в блокчейн (составлен по данным [16])

На рис. 1 представлено взаимодействие участников при обработке транзакций блокчейн.

Процесс обработки транзакций невозможен без информационной среды, которая становится основным условием реализации блокчейн операций.

Последовательность операций осуществляется следующим образом: пользователь отправляет информацию на серверы сети. Майнеры выполняют подбор ключа (хеша), они получают транзакции пользователей от серверов, подбирают блок и отправляют его обратно. Серверы проверяют правильность всех транзакций в блоке, самого блока и добавляют его в блокчейн. После подтверждения транзакции пользователь получает доступ к средствам на его электронном кошельке [16].

Можно выделить два вида блокчейн:

- публичный – база полностью открыта, любой пользователь имеет доступ ко всем транзакциям системы;
- приватный – предполагаются разные уровни доступа к чтению и модификации цепочки.

Технология блокчейн имеет следующие специфические характеристики:

1. Система блокчейн ведет учет обмена данными, каждый обмен данными называется транзакцией, проверенная транзакция добавляется в цепочку в виде блока и не может быть изменена.
2. Децентрализация – нет единого сервера, компьютеры участников обслуживают цепочку вместе.
3. Прозрачность и доступность – информация о транзакциях доступна каждому участнику сети, данные, зафиксированные в блоке цепочки, невозможно изменить.
4. Конфиденциальность – пользователь не может определить получателя и отправителя информации, для проведения транзакций (операций) требуется уникальный ключ доступа.
5. Неограниченность – в теории потенциальный размер цепочки блокчейн бесконечен, на практике он ограничивается вычислительными ресурсами сети.
6. Надежность – для записи нового блока необходимо согласование узлов, отдельный участник (компьютер) не может внести в цепочку недостоверную информацию, изменить ее или фальсифицировать.
7. Безопасность – математико-криптографическая защита информации, позволяющая отклонить внесение несанкционированных изменений вследствие несоответствия предыдущим копиям [4, 17].

С учетом рационального размещения и хранения информации в цепочке блокчейн объем хранения данных может быть уменьшен в 2-3 раза, что приведет к снижению стоимости хранения. Технологии блокчейн позволяют избежать дублирования трафика для почтовых рассылок и доступа к базе, что снизит объем трафика примерно в 4,8 раза. Данная технология включает универсальные интерфейсы формирования данных и доступа к ним, что можно внедрить в любое приложение хранения данных, сократив затраты на разработку и внедрение. Использование стандартизированных интерфейсов способствует снижению расходов на разработку и внедрение приложений хранения данных примерно на 25-30% [9].

Зарубежные авторы выделяют значимую роль блокчейн в развитии цифровой среды, информационного пространства, информационных систем, они отмечают революционный характер данной технологии в сфере экономики, цифровых сервисов, финансовой сферы, организации бизнеса. Исторически появление технологий блокчейн взаимосвязано с развитием криптовалюты и биткоина. Блокчейн затрагивает вопросы цифровой информационной среды на глобальном уровне, что приводит к пересмотру основ финансовых, экономических, информационных систем, некоторых аспектов безопасности [18].

Некоторые авторы исследуют не только саму технологию блокчейн, но и её функции, свойства, причины возникновения, обращая внимание на особенности системы оплаты с использованием блокчейн, внедрение данной технологии в государственные структуры власти. При этом отмечается новый способ передачи и хранения информации с использованием блокчейн, определяется архитектура блокчейн, перспективы развития данной технологии в сфере бухгалтерского учета и аудита [19, 20].

В зарубежных публикациях приводятся примеры использования блокчейн в различных сферах: финансовых транзакциях, построении бизнес-моделей, сервисах оплаты, переходе к новой цифровой валюте, отдельно изучаются правовые аспекты использования блокчейн [21].

СВОЙСТВА ТЕХНОЛОГИИ БЛОКЧЕЙН

Единой точки зрения относительно достоинств и недостатков блокчейн не существует, иногда авторы высказывают противоречивые мнения.

Можно выделить следующие недостатки блокчейн:

1) значительное увеличение цепочки блокчейн. Размер блокчейна криптовалюты биткоин составляет уже больше 200 ГБ. А если количество транзакций (операций) в системе равнялось бы показателю платежной системы Visa или MasterCard, то размер блокчейна был бы намного больше, что может привести к замедлению скорости транзакций (операций);

2) невозможность отмены транзакции (операции). Данное обстоятельство может содействовать мошенничеству: перечислив деньги на счет, вернуть их будет уже невозможно;

3) несмотря на повышенный уровень безопасности данных, отмечается уязвимость блокчейна к возможным кибератакам. Можно нарушить целостность блокчейна, если один участник (злоумышленник) объединит у себя 51% вычислительной мощности сети. На практике такая ситуация практически невозможна для крупных криптовалют;

4) отсутствие или непроработанность законодательно-правовой базы в ряде стран, неясный юридический статус блокчейна;

5) для пользователей криптовалют существуют риски [4, 17].

На рис. 2 в обобщенном виде представлены основные свойства технологии блокчейн.

Технологии блокчейн – универсальный способ хранения и обработки информации практически в любой сфере деятельности, тесно связанный с обеспечением информационной безопасности экономических и других процессов.

Среди современных угроз информационной безопасности увеличивается количество кибератак, случаев кибермошенничества, кибертерроризма, а также совершенствуются методы взлома и нанесения вреда цифровым платформам, появляются новые типы вирусов.

Вопросы информационной безопасности регулируют несколько нормативно-правовых актов:

- Указ Президента РФ от 31.12.2015 г. № 683 «О стратегии национальной безопасности Российской Федерации» (http://base.garant.ru/71296054/);
- Указ Президента РФ от 22.12.2017 г. № 620 «О совершенствовании государственной системы обнаружения, предупреждения и ликвидации последствий компьютерных атак на информационные ресурсы Российской Федерации» (http://base.garant.ru/71840924/?_utl_t=vk);
- Указ Президента РФ от 13.05.2017 г. № 208 «О стратегии экономической безопасности Российской Федерации на период до 2030 года» (http://base.garant.ru/71672608/);
- Распоряжение Правительства РФ от 28.07.2017 г. № 1632-р Об утверждении программы «Цифровая экономика Российской Федерации» (http://base.garant.ru/71734878/);
- Постановление Правительства РФ от 02.03.2019 г. № 234 (ред. от 07.12.2019 г.) «О системе управления реализацией национальной программы «Цифровая экономика Российской Федерации» (http://base.garant.ru/72190034/).

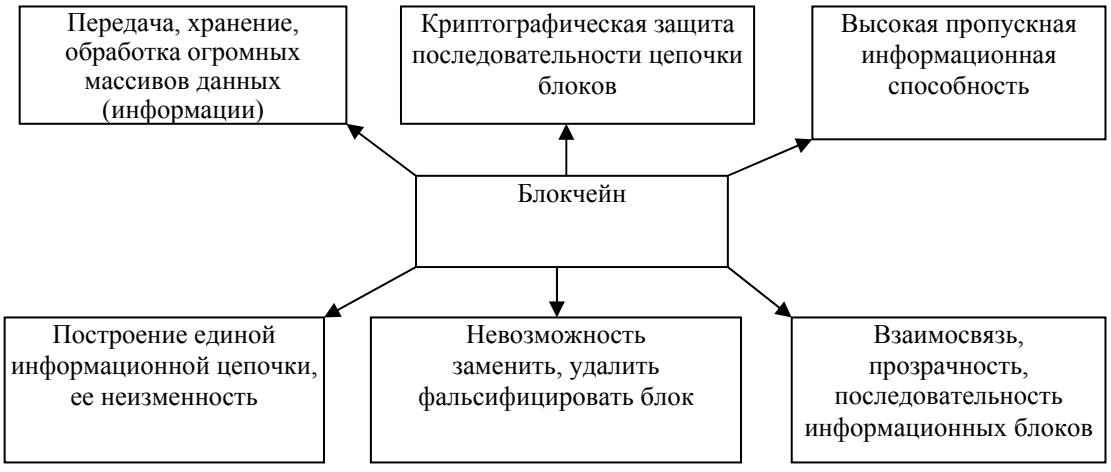

Рис. 2. Основные свойства технологии блокчейн

Таблица 2

Показатели информационного общества*

Показатель	Год				
	2015	2016	2017	2018	2019
Численность студентов, принятых в государственные вузы по направлению «Информатика и вычислительная техника» (человек на 10000 населения)	10	11	12	14	15
Численность выпускников государственных вузов по направлению «Информатика и вычислительная техника» (человек на 10000 населения)	6	7	8	8	8
Удельный вес занятых в сфере информационно-коммуникационных технологий в общей численности занятого населения (%)	1,7	1,7	1,7	1,6	1,7
Количество специалистов по информационно-коммуникационным технологиям (на 10000 работников организаций)	213	218	230	231	234
в том числе: специалистов высшего уровня квалификации (на 10000 работников организаций)	131	136	142	146	149
специалистов среднего уровня квалификации (на 10000 работников организаций)	82	82	88	85	85

* Составлена по материалам [24]

В сфере виртуальных преступлений сегодня выделяются две основные тенденции: (1) в большей степени они направлены на карточный процессинг (обработка платежей по дебетовым и кредитным картам), в меньшей – затрагивают интернет-банкинг (дистанционный доступ к банковским услугам посредством устройства с доступом в Интернет); (2) преступления в криптоиндустрии (криптовалюты, электронные кошельки) увеличиваются [22].

В плане информационной безопасности следует обратиться к понятию незаконный информационный продукт, включающему информацию, содержащую:

- материалы экстремистского, антиправительственного характера, призывающие к смене власти либо компрометирующая деятельность органов власти;
- государственную или военную тайну;
- сведения, способные нанести ущерб или вред экономическим субъектам страны (коммерческая тайна, инсайдерская информация).

Под незаконными информационными услугами мы понимаем сбор, хранение, обработку, реализацию незаконного информационного продукта [23].

Таким образом, технология блокчейн может стать основой кибербезопасности, защиты данных от взлома, несанкционированного доступа или уничтожения информации.

Блокчейн как распределенная база данных обеспечивает новые возможности для хранения и передачи информации, предоставляет новые способы борьбы с киберпреступлениями.

Количественные показатели информационного общества (табл. 2) необходимо проанализировать в аспекте развития технологии блокчейн.

Как видно из табл. 2, за последние пять лет увеличилось количество студентов, принятых в вузы по направлению «Информатика и вычислительная техника», а также численность выпускников, тенденцию к росту имеет и численность специалистов по информационно-коммуникационным технологиям в организациях. В то же время доля занятых в сфере информационно-коммуникационных технологий практически не изменяется. Представленные данные показывают значимость новых информационных технологий в современном обществе, что способствует развитию блокчейн и указывает на перспективность данной технологии.

Вопросы информационной безопасности в контексте блокчейн

Обсуждение технологий блокчейн связывают и с информационной безопасностью общества. В табл. 3 показано, что в современном обществе актуальными являются статистические данные по количеству граждан, которые не используют Интернет по причинам безопасности.

Доля населения, отказывающегося от использования Интернета по причинам безопасности, очень незначительна и мало изменяется год от года. Значительно больше причин отказа связано с отсутствием желания, интереса, навыков работы, технической возможностью подключения.

В табл. 4 приведены угрозы информационной безопасности, с которыми сталкивалось население России, пользующееся Интернетом.

Наиболее распространенная угроза информационной безопасности – это несанкционированная реклама. За рассматриваемый период 2016-2018 гг. отмечается снижение угроз заражения вирусами и несанкционированного доступа к ресурсам, что можно объяснить улучшением навыков работы населения с информационными ресурсами в сети Интернет. Увеличивается и доля населения, которая не сталкивается с угрозами информационной безопасности. Представленные в табл. 4 данные показывают необходимость развития информационных технологий для защиты информации и снижения угроз информационной безопасности в современном цифровом обществе. Блокчейн имеет дальнейшие перспективы развития в данном аспекте.

Таблица 3

Население, не использующее Интернет по причинам безопасности*

Показатель	Год		
	2016	2017	2018
Доля населения, не использующего Интернет по причинам безопасности, в общей численности населения (%)	0,5	0,6	0,4
Доля населения, не использующего Интернет по причинам безопасности, в численности населения, не пользующегося Интернетом (%)	2,2	3,2	2,4
в том числе: ограничение детей от нежелательной информации или программ (%)	0,5	0,4	0,4
защита персонального компьютера от вирусов и вредоносных программ (%)	0,2	0,2	0,1
нежелание раскрывать в Интернете свои персональные данные (%)	1,5	2,6	1,9

* Составлена по материалам [25]

Таблица 4

Угрозы информационной безопасности для населения (% от численности населения, пользующегося Интернетом)*

Угроза	Год		
	2016	2017	2018
Несанкционированная рассылка (спам)	18,4	18,5	19,7
Заражение вирусами	13,4	11,4	8,9
Несанкционированный доступ к информационным ресурсам, системам	3,1	2,7	2,1
Посещение детьми нежелательных сайтов, контакты с потенциально опасными людьми через Интернет	0,8	0,8	0,7
Хищение денежных средств или персональных данных	0,2	0,3	0,2
Другие угрозы	2,9	3,2	2,4
Не сталкивались с угрозами информационной безопасности	61,2	63,1	66

* Составлена по материалам [25]

Таблица 5

Организации, использующие различные средства защиты информации (в процентах от общего числа организаций)*

Средство защиты информации	Год		
	2016	2017	2018
Электронная цифровая подпись	77,2	77,2	78,9
Обновляемые антивирусные программы	76,3	77,2	79,2
Средства аутентификации пользователей	55,6	56,9	60,2
Программные, аппаратные средства, препятствующие несанкционированному доступу	50,8	52,5	56,0
Спам-фильтр	42,7	45,0	48,6
Средства шифрования	42,9	44,3	45,8
Системы обнаружения вторжения в сеть или компьютер	32,8	34,4	37,3
Программные средства контроля защищенности компьютерных систем	26,9	28,1	31,2
Резервное копирование	24,1	24,2	26,5
Биометрические средства аутентификации пользователей	4,1	4,6	6,6

* Составлена по материалам [25]

В табл. 5 представлена доля организаций, использующих различные средства защиты информации в процентах от общего числа организаций.

Данные табл. 5 показывают увеличение использования различных средств защиты информации в организациях. Наряду с такими традиционными средствами как электронная цифровая подпись, антивирусные программы, увеличивается доля организаций, использующих современные средства защиты: спам-фильтр, системы обнаружения вторжения в сеть, программные средства контроля защищенности компьютерных систем, биометрические средства аутентификации. Это свидетельствует о значимости для организаций совершенствования средств защиты информации и данных. В связи с этим развитие и внедрение технологии блокчейн как одного из средств защиты информации является актуальным и перспективным.

БЛОКЧЕЙН В РОССИИ И ЗА РУБЕЖОМ

В августе 2017 г. в России был создан комитет «Программно-аппаратные средства технологии распределенного реестра и блокчейн». Крупные коммерческие банки заинтересованы во внедрении технологий блокчейн. В 2016 г АО Сбербанк России разработал сервис по оформлению защищенных сделок на основе блокчейна. В декабре 2016 г. первая операция в форме аккредитива осуществлена с использованием блокчейна между АО «Альфа-Банк» и авиакомпанией «S7 Airlines». С октября 2017 г. ПАО Сбербанк и АО «Альфа-Банк» используют технологию блокчейн при проведении факторинговых операций. Это позволило включить в данные транзакции большое количество поставщиков при сохранении конфиденциальности проводимых сделок.

Центральным банком РФ на основе блокчейн была разработана технология мастерчейн, которая позволяет проводить финансовые платежи между участниками, а также хранить актуальную информацию о сделках и клиентах и создавать различные финансовые сервисы.

В 2017 г. в Москве технологии блокчейн стали применяться при внесении сведений в Единый государственный реестр недвижимости [2].

До 2024 г. в России ожидается запуск нескольких крупных проектов по внедрению блокчейн в таких сферах, как:

- очереди в садик, школу, получение различных социальных услуг, субсидий;
- автоматизация различных бизнес-процессов (идентификация пользователей, факторинг, банковские гарантии, операции с ценными бумагами);
- образование и здравоохранение – цифровые платформы на основе блокчейн позволят сохранить и обеспечить верификацию «цифрового следа» граждан;
- электронное межведомственное взаимодействие [26].

В зарубежных странах процесс внедрения блокчейн развивается разными темпами. Например, Эстония реализовала проект «Электронная Эстония» с применением технологий блокчейн. Гражданам Эстонии выдаются криптографические защищенные цифровые ID-карты с технологией блокчейн, которые позволяют получать доступ к различным государственным услугам. С помощью этих карт граждане могут видеть информацию о них, хранящуюся в государственных базах данных. Программа «Электронная Эстония» охватывает:

- электронную идентификацию личности (ID-карты, мобильные ID, смарт-ID);
- внутренние процессы (регистр земельных участков, регистр численности населения);
- обеспечение безопасности (электронные подписи, электронный закон, электронная полиция);
- здравоохранение (электронные рецепты на лекарства, контроль состояния здоровья);
- государственные услуги (электронное голосование, правительственное облако, информационное посольство);
- мобильные сервисы (умный транспорт, мобильная парковка);
- ведение бизнеса (электронный налог, электронный банкинг, электронный реестр бизнеса).

Грузия применяет блокчейн в земельном кадастре для подтверждения права собственности на землю и проверки сделок с недвижимостью.

В республике Беларусь использование блокчейн проходит в экспериментальном режиме, есть отдельные проекты в банковской сфере, но пока рано говорить о повсеместном использовании данной технологии.

В Казахстане отсутствует запрет на криптовалюты, однако официального разрешения на них также нет.

В Киргизии блокчейн технологии применяются в деятельности Национального банка.

Армения находится на этапе изучения опыта других стран в данном вопросе, идет разработка правовых аспектов использования блокчейн.

В США блокчейн используется для архивирования и шифрования государственных документов.

В Великобритании технологии блокчейн применяются в сфере социального обеспечения.

В 2019 г. в Индии началась национальная программа по развитию блокчейн. Правительство Индии выделило 900 тыс. долл. для реализации проекта «Распределенный центр передового опыта блокчейн технологий». Технологии блокчейн используются для регистрации собственности, подтверждения подлинности дипломов, сертификатов, договоров и других документов.

Республика Корея модернизирует свою столицу Сеул в соответствии с градостроительным блокчейн-планом на 2018-2022 гг. В Сеуле технологии блокчейн используются в проекте «Умный метрополитен», при минимизации затрат на нерентабельный транспорт, составлении маршрутов для водителей такси с учетом загруженности дороги и времени суток, снижении ДТП с участием детей и людей пожилого возраста.

На Мальте утверждены законы по регулированию технологии блокчейн.

Евразийский экономический союз уделяет особое внимание вопросам использования и регулирования технологий блокчейн, криптоактивов, формированию единой платежной системы [14, 26].

К тестированию и внедрению технологий блокчейн приступили Сингапур, Финляндия, Швейцария, предполагается, что число таких стран ежегодно будет увеличиваться.

ВЫВОДЫ

Перспективы использования блокчейн имеют контур неопределенности, несмотря на инновационный потенциал данной технологии для бизнеса и общества в целом. Центральный банк РФ занимает сдержанную позицию относительно блокчейн, высказывается ряд опасений по возможным рискам нелегальной деятельности в данной сфере.

Ключевую роль в реализации технологии блокчейн играет информация. Все большее количество сфер деятельности информационного общества охватывается технологией блокчейн: бизнес, финансы, государственные услуги, здравоохранение, транспорт и многие другие.

Перспективы развития блокчейн связывают с дальнейшим развитием цифрового информационного общества. Технологии блокчейн будут охватывать новые сферы, например, систему государственного и муниципального управления. По прогнозам планируется проведение переписи населения и выборов с использованием технологий блокчейн.

Одним из важных аспектов для информационного общества в блокчейн является свойство невозможности удаления архивных данных.

Значительный потенциал для блокчейн прогнозируют в управлении инфраструктурой умных городов, в процессах таможенной и транспортной логистики. В промышленной индустрии также существуют широкие перспективы применения блокчейн.

Степень вовлеченности стран в использование данной технологии находится на разных уровнях ее применения и принятия в целом. Исследование данного вопроса показало, что существуют различные точки зрения на перспективы блокчейн, включая как позитивные прогнозы, так и скептические мнения. Данный вопрос требует дальнейшего изучения.

СПИСОК ЛИТЕРАТУРЫ

1. Международная конференция «Цифровая трансформация: интеллектуальная собственность и блокчейн-технологии». – URL: https://rospatent.gov.ru/ru/sourses/multimedia/ blockchain-conference-online (дата обращения: 10.12.2020).
2. Садчиков М.Н., Курбатов Н.М. К вопросу правового регулирования и обеспечения информационной безопасности при использовании технологии блокчейн в банковском секторе экономики Российской Федерации // Вестник Саратовской государственной юридической академии. – 2020. – № 1 (132). – С. 219-229.
3. Степанов А.Е., Серебровский С.П. Исследование направлений успешного внедрения технологии блокчейн в отрасли // Вестник академии. – 2019. – № 4. – С. 16-27.
4. Попов С.А., Охрицкий И.С. Технология блокчейн и его применение сегодня // Инновационные технологии в машиностроении, образовании и экономике. – 2019. – Т. 25, № 4. – С. 53-55.
5. Юшин И.В. Технология «блокчейн» в современной экономике: проблемы и перспективы // Этносоциум и межнациональная культура. – 2018. – № 9(123). – С. 25-36.
6. Антонян Е.А. Вопросы применения новых технологий в противодействии кибертерроризму // Мониторинг правоприменения. – 2020. – № 1(34). – С. 51-55.
7. Леонова Н.Г. Финансовые риски и новые информационные технологии // Наука и бизнес: пути развития. – 2018. – № 3(81). – С. 62-64.
8. Крылов Г.О., Токолов А.В. Влияние блокчейн на мировую экономику // Вестник экономической безопасности. – 2020. – № 1. – С. 192-197.
9. Рязанова А.А. Технология блокчейн в научно-информационной деятельности // Научно-техническая информация. Сер. 1. – 2018. – № 4. – С. 8-12; Ryazanova A.A. The Blockchain Technology in Scientific and Information Activities // Scientific and Technical Information Processing. – 2018. – Vol. 45, № 2. – P. 70-74.
10. Сазанова Е.В. Технология блокчейн в контексте информационной безопасности // Научно-техническое и экономическое сотрудничество стран АТР в XXI веке. – 2019. – Т. 1. – С. 94-97.
11. Амиргамзаев Г.Г., Магомедова Х.А. Использование блокчейн-технологии для хранения учетных данных // Наука: общество, экономика, право. – 2020. – № 2. – С. 268-273.
12. Романенко Н.Ю., Степнова О.В. Технологии блокчейн как процесс развития экономических систем цифровой экономики // Modern economy success. – 2020. – № 1. – С. 232-237.
13. Харченко О.В. Блокчейн в информационном обществе // Вестник Саратовского государственного социально-экономического университета. – 2018. – № 2(71). – С. 28-30.
14. Криптовалюты и блокчейн как атрибуты новой экономики. Разработка регуляторных подходов: международный опыт, практика государств-членов ЕАЭС, перспективы для применения в Евразийском экономическом союзе. – Москва, 2019.
15. Лясников Н.В., Буркальцева Д.Д. Проблемы поддержания работы информационной инфраструктуры в рамках экосистемы цифровой экономики в условиях сбоев при использовании технологии блокчейн // Экономика и социум: современные модели развития. – 2019. – Т. 9, № 2(24). – С. 219-230.
16. Дюдикова Е.И. Блокчейн в национальной платежной системе: сущность, понятие и варианты использования // Инновационное развитие экономики. – 2016. – № 4(34). – С. 139-149.
17. Кузина В.В. Особенности информационной технологии blockchain // Вестник ПГУАС: строительство, наука и образование. – 2019. – № 2(9). – С. 75-78.
18. Ghosh J. The blockchain: opportunities for research in information systems and information technology // Journal of global information technology management. – 2019. – Vol. 22, Issue 4. – P. 235-242.

19. Casino F., Dasaklis T.K., Patsakis C. A systematic literature review of blockchain-based applications: Current status, classification and open issues // Telematics and informatics. – 2019. – Vol. 36. – P. 55-81.
20. Bonson E., Bednarova M. Blockchain and its implications for accounting and auditing // Meditari accountancy research. – 2019. – Vol. 27, № 5. – P. 725-740.
21. Beck R., Avital M., Rossi M., Thatcher J.B. Blockchain technology in business and information systems research // Business & information systems engineering. – 2017. – № 59. – P. 381-384.
22. Борисова Е.С., Белоусов А.Л. Инновации как инструмент обеспечения информационной безопасности и повышения эффективности деятельности банковской системы // Актуальные проблемы экономики и права. – 2019. – Т. 13, № 3. – С. 1330-1342.
23. Лимарев П.В., Лимарева Ю.А. Характеристика рынка информации, используемой в противоправных целях // Научно-техническая информация. Серия 1: Организация и методика информационной работы. – 2018. - № 4. – С. 13-15.
24. Федеральная служба государственной статистики. – URL: https://rosstat.gov.ru/folder/14478 (дата обращения: 10.12.2020)
25. Информационное общество в Российской Федерации. 2019: статистический сборник [Электронный ресурс] / М.А. Сабельникова, Г.И. Ибдрахманова, Л.М. Гохберг, О.Ю. Дудорова и др. Федеральная служба государственной статистики; Национальный исследовательский университет «Высшая школа экономики». – Москва: НИУ ВШЭ, 2019. – ISBN 978-5-7598-2053-6.
26. Истомин Е.П., Кирсанов С.А., Леонтьев Д.В. Некоторые аспекты применения блокчейн-технологий в современной экономике // Информационные технологии и системы: управление, экономика, транспорт, право. – 2020. – № 1(37). – С. 88-102.

Материал поступил в редакцию 14.12.20.

Сведения об авторах

КРЫЛОВА Наталья Павловна – кандидат педагогических наук, доцент кафедры экономики и управления Череповецкого государственного университета, г. Череповец
e-mail: ntlkrylova@rambler.ru

ЛЕВАШОВ Евгений Николаевич – старший преподаватель кафедры экономики и управления Череповецкого государственного университета, г. Череповец
e-mail: levashov_evgenii@mail.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.4’414:81’373.424

А.Ю. Малафеев, Л.П. Мальтина

Распределённые представления редких слов русского языка, учитывающие векторы однокоренных слов*

Рассматриваются алгоритмы, выполняющие автоматический морфемный анализ слов, и методы распределённых представлений слов, которые используют информацию о морфемном составе, но не напрямую, а через усреднение векторов однокоренных слов. Оценивается качество моделей морфемного анализа для русского языка, в том числе и на выборке из редких слов. Предлагается несколько способов получения распределённых представлений редких слов на основе word2vec-представлений однокоренных слов. Проведённые эксперименты показали, что на задаче определения семантической близости пары слов предлагаемые методики дают результаты, сопоставимые с результатами модели fastText или превосходят их.

Ключевые слова: *распределённые представления слов, word2vec, fastText, морфемный анализ, однокоренные слова, редкие слова, семантическая близость пары слов*

DOI: 10.36535/0548-0027-2021-02-2

ВВЕДЕНИЕ

В настоящее время в задачах, связанных с автоматической обработкой текстов на естественном языке, применяются методы построения распределённых представлений слов для моделирования лексической семантики. Эти методы позволяют отображать слова в том или ином языке в Евклидово пространство относительно малой размерности $\mathbb{R}^d$, где количество слов обычно не превышает нескольких сотен.

Одна из наиболее известных и широко применяемых компьютерных дистрибутивных моделей лексической семантики – word2vec [1], основной принцип работы которой заключается в том, что распределённые представления слов являются результатом обучения искусственной нейронной сети на некотором корпусе текстов. Но если частота слова в обучающем корпусе ниже установленного порога, модель word2vec не может предоставить информацию о семантике этого слова. С этой проблемой справляется модель fastText [2], так как учитывает не только слова, но и их подстроки, *n*-граммы определённого диапазона. Однако *n*-граммы далеко не всегда соответствуют морфам, т. е. реализациям минимальных значимых единиц языка. Например, если модель fastText учитывает n-граммы длиной от четырёх до пяти символов, то для слова *шумозаградительный* будут рассматриваться 29 *n*-грамм. Из них только две (*-град-* и *-тельн-*) совпадают с морфами, действительно присутствующими в слове, а ещё пять морфов, которые есть в этом слове (*-шум-, -о-, -за-, -и- , -ый-*), отсутствуют в списке *n*-грамм. Это свидетельствует о том, что, возможно, часть информации, содержащейся в модели fastText, – избыточна. Действительно, модели fastText занимают больше памяти, чем модели word2vec. Если сравнить, например, модели, взятые с сайта (https://rusvectores.org) [3] и обученные на Национальном корпусе русского языка (НКРЯ), то размер модели fastTest (http://vectors.nlpl.eu/repository/20/181.zip) почти в пять раз больше, чем модели word2vec (http://vectors.nlpl.eu/repository/20/180.zip) – 2,4 Гб и 462 Мбайт, соответственно. Кроме того, для внесловарных слов fastText больше ориентируется на орфографическое сходство, чем на семантическое, поэтому иногда возможны ошибки. Например, для модели fastText, обученной на НКРЯ, в десятку наиболее близких для слова *раннеперестроечный* попадают слова *троечный* и *троечник*.

Решить проблему внесловарных слов, в большей степени ориентируясь на их семантику, чем на орфографию, и при этом не слишком увеличивать размер модели, позволяет учёт морфемного состава слов. Для выполнения морфемного анализа, можно было бы использовать словарь, но для большого количества слов готовые морфемные разборы отсутствуют. Например, в НКРЯ более миллиона словоформ, которые встречаются не менее трёх раз, в то время как

* Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2019-2020 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

в самом большом находящемся в открытом доступе морфемном словаре для русского языка [4], менее ста тысяч разборов. Это связано с тем, что в словаре приводятся разборы только для начальной формы слова и отсутствуют многие неологизмы и термины, и поэтому создание распределённых представлений, которые учитывают морфемный состав слов, происходит в два этапа: сначала выполняется автоматический морфемный анализ, затем применяется тот или иной алгоритм, представляющий слова в виде векторов относительно небольшой размерности (например, 300), используя информацию об их морфемном составе.

В настоящей работе акцент делается на втором этапе. Так как основную лексическую информацию в слове содержит корень, предлагается дополнить обычную модель word2vec распределёнными представлениями внесловарных слов, полученными на основе word2vec-векторов их однокоренных слов. Качество дополненной модели оценивается на задаче определения семантической близости пары слов. Разработанный нами способ был применён для трёх моделей word2vec, обученных на разных корпусах, и на всех из них показал результаты, близкие к результатам fastText или превосходящие их[1].

ОБЗОР МЕТОДОВ МОРФЕМНОГО АНАЛИЗА И СОЗДАНИЯ РАСПРЕДЕЛЕННЫХ ПРЕДСТАВЛЕНИЙ

Существует множество алгоритмов морфемного анализа: на основе униграммной вероятностной модели [5, 6]; условных случайных полей [7, 8]; *sequence to sequence* нейронной сети [9]; свёрточных нейронных сетей [10]; градиентного бустинга на деревьях решений [11]; рекуррентных нейронных сетей с двунаправленной долгой краткосрочной памятью [12].

Что касается создания распределённых представлений, учитывающих морфемный состав слов, здесь можно выделить три основных подхода.

Первый заключается в том, что для слов, которые есть в словаре модели и векторы слов остаются прежними, а для внесловарных слов берутся векторы, наиболее близких им по морфемному составу. Этот подход применялся в работе [9], где в качестве вектора, представляющего внесловарное слово, использовался вектор для части этого слова, которая присутствовала в словаре модели word2vec. Такой частью может быть один морф или несколько подряд идущих морфов. Например, если слово *жертвовательница* является внесловарным и разбивается на морфемы как *жертв/ова/тель/ниц/а*, то для него берётся тот же вектор, что и для слова *жертвователь*, найденного в словаре модели. Это позволяет повысить качество определения семантической близости для пары слов.

При втором подходе вместо слов или *n*-грамм при обучении моделей word2vec и fastText используются морфы. В [13] исследователи описывают обучение модели word2vec на морфах, но на выборке для определения семантической близости пар редких многоморфемных слов эти модели показывают результаты ниже, чем fastText. В работе [14] стандартная модель fastText модифицирована так, что вместо *n*-грамм она основывается на морфах. Используя внешние и внутренние способы оценивания распределённых представлений, авторы [14] демонстрируют, что морфемный fastText обычно работает лучше, чем SkipGram (разновидность модели word2vec), но хуже, чем классический fastText.

В третьем подходе представления слов конструируются из векторов их морфов. В [15] морфемам в соответствие ставятся их дефиниции. Для каждого морфа в слове из набора дефиниций выбирается та, вектор word2vec которой ближе к word2vec-вектору слова. Итоговый вектор слова складывается из его word2vec-вектора и векторов его морфов, взятых с весами, пропорциональными вкладу морфа в значение слова. Исследователи [16] получают вектор слова, минимизируя функцию потерь, характеризующую различие между предобученным word2vec-вектором и новым вектором слова, который является взвешенной суммой векторов для нескольких вариантов морфемного анализа слова; веса определяются с помощью механизма внимания. Вектор для одного из вариантов морфемного анализа представляет функцию от векторов морфов в составе слова, вычисляемую с помощью рекуррентной нейронной сети с двунаправленной долгой краткосрочной памятью. Такие распределённые представления, основанные на информации о морфемном составе слова, оказываются лучше векторов на основе слов, символов или *n*-грамм для языков со сложной морфологией.

Рассматривая новизну предлагаемого нами метода создания распределённых представлений для редких слов и его отличия от существующих методов, следует сначала отметим, что наш метод близок к первому подходу, так как он применяется только для редких слов, векторы для которых строятся на основе векторов уже известных слов. Отличия нашего метода от описанного в [9] в том, что, во-первых, в качестве вектора редкого слова выбирается не готовый вектор, а взвешенное усреднение нескольких векторов (логично предположить, что если слова не совпадают, хотя и являются близкими по морфемному составу, они должны представляться неодинаковыми векторами), во-вторых, слова, векторы которых усредняются, обязательно должны быть однокоренными для рассматриваемого редкого слова, так как именно корень содержит основную информацию о значении слова.

ИСПОЛЬЗУЕМЫЕ МОДЕЛИ МОРФЕМНОГО АНАЛИЗА

Для морфемного анализа в настоящей работе применяются две модели: униграммная вероятностная модель Morfessor [5, 6] (частичное обучение) и модель, использующая свёрточные нейронные сети [10] (обучение с учителем), выбор которых объясняется тем, что эти модели отличаются по качеству морфемного анализа. Morfessor – одна из первых и относительно простых моделей морфемного анализа, а нейросетевая модель, описанная в [10], в настоящее время показывает один из лучших результатов для русского языка. Это даёт возможность определить, насколько

[1] Написанный нами код на языке программирования *Python*, реализующий эксперименты по оценке моделей, доступен по ссылке: https://github.com/lpmaltina/MorphemicEmbeddings.

важна точность автоматического морфемного анализа для создания распределённых представлений, «обогащённых» информацией об однокоренных словах. В качестве неразмеченных обучающих данных для модели Morfessor был использован список униграмм (http://www.ruscorpora.ru/new/ngrams/1grams-3.zip) из Национального корпуса русского языка (1054210 словоформ). Среди них были отобраны только те словоформы, которые содержали буквы русского алфавита и дефисы. Словоформы были приведены к нижнему регистру, а незнаменательные части речи исключены. Объём выборки составил 674940 словоформ. Лемматизированная версия этой выборки, в которой частоты для форм одних и тех же слов суммировались, включала 146907 лемм. В качестве размеченных данных для обеих моделей использовались разборы слов из морфемно-орфографического словаря А.Н. Тихонова [4], которые были случайным образом разбиты на обучающую, валидационную и тестовую выборки в соотношении 40/30/30 (38368/28777/28777 примеров). Гиперпараметры моделей настраивались на валидационной выборке. В табл. 1 указаны их подобранные значения.

Модель Morfessor выполняет сегментацию слова на морфы, но не определяет тип морфов, т. е. не предоставляет информацию о том, является ли определенный морф корнем, префиксом, суффиксом и т.д. Модель, использующая свёрточные нейронные сети, не только выполняет сегментацию слова на морфы, но и указывает для каждого морфа его тип. Чтобы строить векторы внесловарных слов в рамках предлагаемого нами подхода, требуется находить их корни, поэтому для модели Morfessor используется следующее правило: морф считается корнем, если он не входит в списки аффиксов, взятые из [17]. Так как в этих списках приводится только по одному алломорфу для каждой морфемы, другие возможные алломорфы были добавлены вручную (например, для префикса *из-* были добавлены алломорфы *ис-*, *изо-*).

Таблица 1

Значения гиперпараметров, подобранные для моделей морфемного анализа

Модель морфемного анализа	Гиперпараметр	Значение гиперпараметра
Morfessor	Лемматизация неразмеченной обучающей выборки	Нет
	Учет частотности словоформ	Нет
	α (гиперпараметр для неразмеченных данных, его низкое значение даёт преимущество словарям с короткими морфами, а высокое значение – с более длинными морфами)	0,1
	β (аналогичный гиперпараметр для размеченных данных)	1000
Свёрточные нейронные сети	Свёрточные слои	4
	Ширина окна	5
	Фильтры	240
	Полносвязанные нейроны в предпоследнем слое	64
	Dropout rate	0,4
	Количество моделей в ансамбле	3
	Меморизация – выполняется ли проверка того, что морф, полученный в результате морфемного анализа, может встречаться в данной позиции	Да

Таблица 2

Качество моделей морфемного анализа

Модель морфемного анализа	Точность	Полнота	F1-мера	Доля слов, в которых все морфемные границы проведены верно	Доля слов, в которых все корни выделены верно
Качество на тестовой выборке из словаря А.Н. Тихонова					
Morfessor	0,9085	0,8994	0,9039	0,6930	0,6902
Свёрточные нейронные сети	**0,9645**	**0,9676**	**0,9660**	**0,8521**	**0,8617**
Качество на выборке из редких слов					
Morfessor	0,5027	0,8132	0,6213	0,1750	0,155
Свёрточные нейронные сети	**0,7981**	**0,8298**	**0,8136**	**0,5463**	**0,5438**

Качество морфемного анализа и определения корня в слове оценивалось на тестовой выборке из словаря [4] и на выборке из 800 слов с незнакомыми корнями [18], при использовании которой задача морфемного анализа ставится в усложнённом варианте: слова, которые входят в эту выборку, специально были отобраны так, чтобы хотя бы один из корней отсутствовал в обучающей выборке на основе словаря [4]. Выборка включает неологизмы (*буккроссинг*), термины (*аденозинтрифосфорный*), сленг (*загуглиться*) и производные слова от имён собственных (*неогумбольдтианство*) (табл. 2). Модель, основанная на свёрточных нейронных сетях, показывает более высокие результаты, чем Morfessor. Применение моделей, дающих разное качество морфемного анализа, позволит понять, насколько они влияют на качество распределённых представлений, косвенно использующих информацию о морфемном составе слов.

ПРЕДЛАГАЕМЫЙ МЕТОД ПРЕДСТАВЛЕНИЯ ВНЕСЛОВАРНЫХ СЛОВ

Для получения распределённых представлений по нашему методу сначала проводится морфемный анализ словаря стандартной модели word2vec и внесловарных слов из набора данных для оценки семантической близости. Затем для внесловарных слов строятся векторы на основе однокоренных слов, которые присутствуют в словаре word2vec-модели. Для известных слов, т.е. тех, которые содержатся в словаре стандартной модели word2vec, векторы не меняются. Далее полученная модель применяется для определения семантической близости пар слов.

Перечислим все применённые способы построения распределённых представлений внесловарных слов:

1) *word2vec + усреднение* – в качестве вектора для внесловарного слова используется усреднение векторов для всех слов, в которых имеется хотя бы один корень, что и в этом внесловарном слове;

2) *word2vec + частотные веса* – вектор v для внесловарного слова w_{oov} конструируется из векторов всех однокоренных слов, взятых с частотными весами (f – частота):

$$v(w_{oov}) = \frac{\sum_{i=1}^{N} v(w_i) \cdot f(w_i)}{\sum_{i=1}^{N} f(w_i)};$$

где $\{w_1, ..., w_N\}$ – множество однокоренных слов для слова w_{oov}. Частотные веса для слов берутся из библиотеки wordfreq [19];

3) *word2vec + веса-вероятности* – вектор v для внесловарного слова w_{oov} конструируется из векторов всех однокоренных слов, взятых с весами-вероятностями того, что в этом однокоренном слове был верно выделен корень (p_{root} – такая вероятность):

$$v(w_{oov}) = \frac{\sum_{i=1}^{N} v(w_i) \cdot p_{root}(w_i)}{\sum_{i=1}^{N} p_{root}(w_i)}.$$

Веса-вероятности получены с помощью модели, выполняющей морфемный анализ слов;

4) *word2vec + усреднение наиболее частотных* – способ отличается от способа (1) тем, что для каждого корня, входящего во внесловарное слово, учитываются только первые пять слов с наибольшей частотой. Мотивацией для такого способа получения векторов послужило то, что встречаются высокочастотные корни, для которых есть более сотни однокоренных слов, и усреднение очень большого количества векторов может отрицательно сказаться на результатах;

5) *word2vec + усреднение наиболее вероятных* – отличается от способа (4) тем, что вместо частотных весов используются веса-вероятности того, что корень в слове был выделен верно;

6) *word2vec + усреднение с наибольшей морфемной F1-мерой и частотой* – в основе этого способа лежит предположение, что следует отдавать предпочтение однокоренным словам, у которых большее количество морфов совпадает с морфами слова, для которого строится вектор. Поэтому используется морфемная F1-мера, характеризующая, насколько совпадают морфы внесловарного слова и конкретного однокоренного слова. Для каждого корня однокоренные слова сортируются по убыванию их морфемной F1-меры и частоты. Затем от каждого корня выбирается по три однокоренных слова, их векторы усредняются. Здесь берутся уже три слова, а не пять, так как предполагается, что выбранные слова являются наиболее близкими по своему морфемному составу для слова, для которого строится вектор, а, следовательно, и более семантически близкими.

В морфемной F1-мере в качестве истинно положительных значений *tp* используется количество морфов, совпадающих во внесловарном слове w_{oov} и однокоренном для него слове w_d, с весами, определяющимися в зависимости от типа морфов: количество корней *r* берётся с весом 1, количество префиксов *pref*, суффиксов *suf* и постфиксов *post* – с весом $\gamma = 0{,}3$ (значение параметра подобрано вручную), окончания и соединительные гласные не учитываются. Такие веса были выбраны потому, что корень выражает основное лексическое значение, а префикс, суффикс и постфикс – его дополнительные компоненты. Что касается соединительных гласных, то существование у них какого-либо значения является спорным, иногда им приписывают соединительное значение как у союза *и*. Окончания же выражают только словоизменительное значение. Таким образом:

$$tp = |r(w_d) \cap r(w_{oov})| + \gamma \left(\begin{array}{l} |pref(w_d) \cap pref(w_{oov})| + \\ + |suf(w_d) \cap suf(w_{oov})| + |post(w_d) \cap post(w_{oov})| \end{array} \right)$$

В качестве ложноположительных значений *fp* в морфемной F1-мере выступает количество морфов, которые есть в однокоренном слове, но отсутствуют во внесловарном слове. Веса берутся аналогично:

$$fp = |r(w_{oov}) \setminus r(w_d)| + \gamma \left(\begin{array}{l} |pref(w_{oov}) \setminus pref(w_d)| + \\ + |suf(w_{oov}) \setminus suf(w_d)| + |post(w_{oov}) \setminus post(w_d)| \end{array} \right).$$

Истинно отрицательные значения *tn* в морфемной F1-мере – это количество морфов, которые есть во внесловарном слове, но отсутствуют в однокоренном для него:

$$tn = \left|r\left(w_d\right) \setminus r\left(w_{oov}\right)\right| + \\ +\gamma\left(\begin{array}{l}\left|pref\left(w_d\right) \setminus pref\left(w_{oov}\right)\right| + \\ +\left|suf\left(w_d\right) \setminus suf\left(w_{oov}\right)\right| + \left|post\left(w_d\right) \setminus post\left(w_{oov}\right)\right|\end{array}\right);$$

7) *word2vec + усреднение с наибольшей морфемной F1-мерой и вероятностью* – способ аналогичен предыдущему и отличается от него тем, что для каждого корня однокоренные слова сортируются по убыванию их морфемной F1-меры и вероятности того, что в данном слове был верно выделен корень.

При способах получения распределённых представлений (1), (2) и (5) в качестве моделей морфемного анализа используются Morfessor и модель, основанная на свёрточных нейронных сетях. Так как способы (3), (4), (6) и (7) требуют знания о типе аффиксов (префикс, суффикс, соединительная гласная, окончание, постфикс) и/или вероятности, что морф выделен верно (а модель Morfessor не предоставляет такой информации), то для них применяется только модель, использующая свёрточные нейронные сети.

Если модель, строящая вектор слова на основе однокоренных слов, не выделит корень или не найдёт однокоренных слов с этим корнем, семантическая близость пары слов, в которую входит это слово, считается равной нулю. Отметим, что доля пар с внесловарными словами для моделей, использующих информацию о морфемном составе слова, намного меньше доли пар с внесловарными словами в стандартных word2vec моделях.

ЭКСПЕРИМЕНТ ПО ОЦЕНКЕ КАЧЕСТВА РАСПРЕДЕЛЁННЫХ ПРЕДСТАВЛЕНИЙ

Качество получаемых распределённых представлений, учитывающих векторы однокоренных слов, оценивалось на задаче определения семантической близости для пары слов. Использовался набор данных, состоящий из 104 пар редких многоморфемных слов [13]. Дополнительные эксперименты проводились на тестовых выборках с соревнования RUSSE-2015 в рамках конференции «Диалог» [20]:

- выборки AE (1952 пары слов) и AE2 (3002 пары слов) – для каждой пары указано, есть ли между этими словами ассоциативная связь;
- выборка HJ (333 пары слов) – для каждой пары приведены экспертные оценки семантической близости слов;
- выборка RT (9548 пар слов) – для каждой пары дано, есть ли между этими словами синонимические или гипо-гиперонимические отношения. Частеречные теги для всех слов были добавлены с помощью библиотеки deeppavlov[2].

Распределённые представления, полученные по предлагаемому методу, сравнивались со следующими моделями:

- *стандартная модель word2vec* – в ней отсутствуют векторы для внесловарных слов. Для каждой пары, в которой присутствует внесловарное слово, считается, что семантическая близость этих слов равна нулю;
- *word2vec + случайные векторы* – основой модели является стандартная word2vec модель, но в качестве представлений для внесловарных слов выбираются векторы соответствующей размерности со случайными значениями из полуинтервала [-1; 1) (такой выбор обусловлен тем, что обычно значения векторов word2vec-моделей находятся в этом диапазоне);
- *word2vec + случайное значение семантической близости* – используется стандартная word2vec модель, но для пар, в которых хотя бы одно слово является внесловарным, значение семантической близости определяется случайным образом, оно выбирается из полуинтервала [0, 1);
- *fastText* – стандартная модель fastText.

Эксперимент проводился на распределённых представлениях, обученных на трёх корпусах: НКРЯ, веб-корпусах Araneum [21] и Тайга [22]. Использовались предобученные модели word2vec[3,4,5] и fastText[6,7,8] проекта RusVectores [3].

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ОЦЕНКЕ МОДЕЛЕЙ

В табл. 4 приведены результаты определения семантической близости слов, полученные при помощи распределённых представлений на основе однокоренных слов на наборе данных [13].

Корреляции с экспертными оценками у предложенных нами модификаций word2vec для всех корпусов были выше, чем у стандартной модели. Для двух корпусов (Araneum и Тайга) корреляции предложенных моделей с экспертными оценками оказались выше, чем у fastText модели (для Araneum – 0,7149 и 0,6498; для Тайги – 0,7752 и 0,6615). Для НКРЯ модель fastText имеет корреляцию с экспертными оценками лишь на 0,0023 выше, чем у одной из предложенных моделей, которая строит векторы для внесловарных слов на основе однокоренных (для нашего подхода корреляции равны 0,7242; для модели fastText – 0,7265). Отметим, что корреляции с экспертными оценками обычно выше у моделей, которые ограничены по количеству усредняемых слов от каждого корня. Там, где было возможно сопоставление моделей морфемного анализа Morfessor и модели, использующей свёрточные нейронные сети, в большинстве случаев результаты модели со свёрточными нейронными сетями были несколько выше. Лучшие результаты по определению семантической близости также были получены с помощью модели, использующей свёрточные нейронные сети.

[2] https://github.com/deepmipt/DeepPavlov

[3] http://vectors.nlpl.eu/repository/20/180.zip

[4] https://rusvectores.org/static/models/rusvectores4/Araneum/Araneum_upos_skipgram_300_2_2018.vec.gz

[5] http://vectors.nlpl.eu/repository/20/185.zip

[6] http://vectors.nlpl.eu/repository/20/181.zip

[7] https://rusvectores.org/static/models/rusvectores4/fastText/Araneum_none_fastTextcbow_300_5_2018.tgz

[8] http://vectors.nlpl.eu/repository/20/187.zip

На тестовых выборках соревнования RUSSE-2015 (AE, AE2, HJ, RT) результаты были схожими. Все распределённые представления, учитывающие векторы однокоренных слов для внесловарных слов, демонстрируют более высокое качество, чем стандартные модели word2vec. Хотя при обучении на НКРЯ предложенные модели слегка проигрывают fastText, их результаты сопоставимы. При обучении на корпусе Araneum эти модели показывают более высокое качество, чем модель fastText на трех выборках (AE, AE2, HJ) из четырёх. Обучаясь на корпусе Тайга, модификации word2vec превосходят на всех четырёх выборках модель fastText.

На трёх (AE2, HJ, RT) из четырёх выборок лучше всего проявили себя модели, которые имеют ограничение по количеству усредняемых слов от корня, особенно модели, учитывающие морфемную F1-меру. На этих же выборках разбиение слов на морфемы с помощью свёрточных нейронных сетей оказалось более эффективно, чем с помощью модели Morfessor.

Примеры наиболее семантически близких слов для нескольких редких многоморфемных слов, которые отсутствуют в стандартной модели word2vec, приведены в табл. 5. Результаты показаны для лучшего способа получения векторов для внесловарных слов по каждому корпусу.

Одна из основных причин ошибок моделей – неразличение омонимичных корней. В табл. 6 представлены примеры таких ошибок.

Таким образом, можно сделать вывод, что предложенные модели справляются с задачами определения семантической близости пары слов, а также с похожими задачами (определением того, есть ли между словами ассоциативная связь, или синонимические и гипогиперонимические отношения). Результаты предложенных модификаций word2vec, использующих векторы однокоренных слов для внесловарных слов, превосходят стандартные модели word2vec и fastText. Как правило, качество распределённых представлений слов, полученных в результате применения нейросетевой модели морфемного анализа, несколько выше, чем у моделей, использующих результаты системы морфемного анализа Morfessor. Одной из основных причин ошибок предложенных моделей является неразличение омонимичных морфов.

Таблица 4

Результаты определения семантической близости слов

Способ получения распределённых представлений	Модель морфемного анализа	Корреляция Спирмана с экспертными оценками		
		доля пар с внесловарными словами		
		НКРЯ	Araneum	Тайга
Стандартная модель *word2vec*	Нет	0,3402	0,4179	0,2476
		0,4712	0,2885	0,4808
word2vec + случайные векторы	Нет	0,4386	0,4990	0,3751
		0,0000	0,0000	0,0000
word2vec + случайное значение семантической близости	Нет	0,4574	0,4987	0,3464
		0,0000	0,0000	0,0000
fastText	Нет	**0,7265**	0,6498	0,6615
		0,0000	0,0000	0,0000
word2vec + усреднение	Morfessor	0,6380	0,6818	0,6419
		0,0000	0,0000	0,0000
	Свёрточные нейронные сети	0,6820	0,6873	0,6796
		0,0385	0,0288	0,0385
word2vec + частотные веса	Morfessor	0,6205	0,6581	0,6484
		0,0000	0,0000	0,0000
	Свёрточные нейронные сети	0,6215	0,7061	0.6892
		0,0385	0,0288	0,0385
word2vec + веса-вероятности	Свёрточные нейронные сети	0,6810	0,6827	0,6844
		0,0385	0,0288	0,0385
word2vec + усреднение наиболее частотных	Morfessor	0,6903	0,7066	0,7059
		0,0000	0,0000	0,0000
	Свёрточные нейронные сети	0,6899	**0,7149**	0,7202
		0,0385	0,0288	0,0385
word2vec + усреднение наиболее вероятных	Свёрточные нейронные сети	0,6610	0,6977	0,7272
		0,0385	0,0288	0,0385
word2vec + усреднение с наибольшей морфемной F1-мерой и частотой	Свёрточные нейронные сети	0,7242	0,7139	0,7681
		0,0385	0,0288	0,0385
word2vec + усреднение с наибольшей морфемной F1-мерой и вероятностью	Свёрточные нейронные сети	0,7180	0,7121	**0,7752**
		0,0385	0,0288	0,0385

Таблица 5

Примеры наиболее семантически близких слов для внесловарных слов

Корпус, способ получения распределённого представления и модель морфемного анализа	Слово	10 слов, наиболее семантически близких к данному	
Корпус: НКРЯ **Способ:** *word2vec + усреднение с наибольшей F1-мерой и частотой* **Модель:** свёрточные нейронные сети	восемьсотпятидесятилетие_NOUN	десятилетие NOUN	0,6937
		восемьдесят VERB	0,6506
		восемьдесят NUM	0,6447
		семьдесят NUM	0,6405
		девяносто NUM	0,6146
		шестьдесят NUM	0,6045
		девяносто ADV	0,5965
		двадцать_NUM	0,5785
		тридцать NUM	0,5724
		пятьдесят NUM	0,5632
	древневерхненемецкий_ADJ	верхний ADJ	0,697
		древний ADJ	0,6454
		верхний NOUN	0,6367
		нижний ADJ	0,6241
		древнейший ADJ	0,552
		древний NOUN	0,544
		древность NOUN	0,5222
		шумерский ADJ	0,5159
		древнеегипетский ADJ	0,4984
		сарматский_ADJ	0,4902
	камнесамоцветный_ADJ	самоцвет NOUN	0,875
		самоцветный ADJ	0,8375
		изумруд NOUN	0,6967
		малахит NOUN	0,6957
		яшма NOUN	0,6905
		агат NOUN	0,6845
		топаз NOUN	0,678
		аметист NOUN	0,6738
		сапфир NOUN	0,6721
		яхонт NOUN	0,6681
	раннеперестроечный_ADJ	перестроечный ADJ	0,8977
		постперестроечный ADJ	0,7586
		доперестроечный ADJ	0,7444
		брежневский ADJ	0,7308
		послесталинский ADJ	0,6621
		хрущевский ADJ	0,6549
		позднесоветский ADJ	0,6386
		горбачевский ADJ	0,6345
		застойный ADJ	0,6204
		постсоветский ADJ	0,5881
	шумозаградительный_ADJ	зашуметь VERB	0,7151
		загудеть VERB	0,5939
		зарокотать VERB	0,5769
		загрохотать VERB	0,5573
		заколыхаться VERB	0,529
		загреметь_VERB	0,527
		смолкнуть VERB	0,5243
		забурлить VERB	0,5242
		задвигаться VERB	0,5147
		затрещать VERB	0,5098
	биокибернетик_NOUN	кибернетика_NOUN	0,8522
		биохимик NOUN	0,8112
		биолог NOUN	0,7429
		генетик NOUN	0,7105
		физик NOUN	0,6872
		ученый NOUN	0,6802
		генетика NOUN	0,6741
		биология_NOUN	0,6618
		биохимия NOUN	0,6525
		нейрофизиология_NOUN	0,6458

Корпус: Araneum **Способ:** *word2vec + усреднение наиболее частотных* **Модель:** свёрточные нейронные сети	восемьсотпятидесятилетие_NOUN	шестьдесят_NUM	0,8989
		восемьдесят_NUM	0,8797
		тридцать_NUM	0,8794
		семьдесят_NUM	0,8782
		пятьдесят_NUM	0,8711
		сорок_NUM	0,871
		двадцать_NUM	0,8566
		двести_NUM	0,8539
		триста_NUM	0,8531
		четыреста_NUM	0,8518
	древневерхненемецкий_ADJ	древние_NOUN	0,6767
		древний_ADJ	0,6636
		древность_NOUN	0,651
		этруск_NOUN	0,6172
		древнеславянский_ADJ	0,612
		древнеегипетский_ADJ	0,609
		древнегреческий_ADJ	0,6061
		древнерусский_ADJ	0,6002
		шумерский_ADJ	0,5962
		дохристианский_ADJ	0,588
	камнесамоцветный_ADJ	цветок_NOUN	0,6572
		многоцветный_ADJ	0,5702
		шумозаградительный_ADJ	0,5682
		цвета_NOUN	0,5656
		полудрагоценный_ADJ	0,5539
		цветочек_NOUN	0,5538
		разноцветный_ADJ	0,5518
		древневерхненемецкий_ADJ	0,5505
		цвет_NOUN	0,5488
		хризантема_NOUN	0,5474
	раннеперестроечный_ADJ	постройка_NOUN	0,6644
		построить_VERB	0,6377
		строить_VERB	0,6112
		возводить_VERB	0,6051
		строительство_NOUN	0,5904
		ранна_NOUN	0,5841
		сооружение_NOUN	0,5835
		возведение_NOUN	0,5834
		раннесредневековый_ADJ	0,5805
		сооружать_VERB	0,5782
	шумозаградительный_ADJ	шум_NOUN	0,6517
		шуметь_VERB	0,6453
		грохотать_VERB	0,5856
		валяносапожник_NOUN	0,5805
		грохот_NOUN	0,5805
		греметь_VERB	0,5804
		гул_NOUN	0,5783
		гудеть_VERB	0,5764
		камнесамоцветный_ADJ	0,5682
		канонада_NOUN	0,5681
	биокибернетик_NOUN	кибернетика_NOUN	0,7808
		кибернетик_NOUN	0,7341
		биофизика_NOUN	0,7292
		биолог_NOUN	0,7267
		биология_NOUN	0,71
		генетика_NOUN	0,688
		биоинформатика_NOUN	0,6789
		биологический_ADJ	0,6755
		зоология_NOUN	0,6676
		нейрофизиология_NOUN	0,6637

Корпус: Тайга **Способ:** *word2vec + усреднение с наибольшей F1-мерой и вероятностью* **Модель:** свёрточные нейронные сети	восемьсотпятидесятилетие_NOUN	семьдесят_NUM	0,755
		двадцать_NUM	0,7489
		восемьдесят_NUM	0,747
		тридцать_NUM	0,7379
		восемьдесят_VERB	0,7306
		шестьдесят_NUM	0,7227
		пятьдесят_NUM	0,7193
		сорок_NUM	0,7008
		девяносто_NUM	0.6999
		семьдесят_VERB	0,686
	древневерхненемецкий_ADJ	верхний_NOUN	0,663
		верхний_ADJ	0,6588
		нижний_ADJ	0,6314
		древний_ADJ	0,6224
		древневосточный_ADJ	0,6219
		древнейший_ADJ	0,6076
		верхний_PROPN	0,6065
		древний_NOUN	0,5988
		-окский_ADJ	0,5961
		нижний_NOUN	0,5907
	камнесамоцветный_ADJ	самоцветный _ADJ	0,7568
		самоцвет_NOUN	0,746
		камень_NOUN	0,7078
		каменья_NOUN	0,6878
		камни_NOUN	0,6602
		камння_NOUN	0,6335
		камушек_NOUN	0,6234
		изумруд_NOUN	0,6223
		камешек_NOUN	0,608
		алмаз_NOUN	0,6062
	раннеперестроечный_ADJ	перестроечный_ADJ	0,7896
		постперестроечный_ADJ	0,781
		постсоветский_ADJ	0,7532
		перестроечной_ADJ	0,7283
		позднесоветский_ADJ	0,7277
		раннесоветский_ADJ	0,7079
		послереволюционный_ADJ	0,676
		послесталинский_ADJ	0,6498
		горбачевский_ADJ	0,644
		послесоветский_ADJ	0,6427
	шумозаградительный_ADJ	зашуметь_ADV	0,6651
		загудеть_VERB	0,6142
		загудеть_NOUN	0,5652
		зашуметь_VERB	0,6732
		заградитель_NOUN	0,5564
		загрохотать_VERB	0,5533
		шумить_VERB	0,5468
		шуметь_VERB	0,5437
		засвистеть_VERB	0,5418
		зашуметь_NOUN	0,5409
	биокибернетик_NOUN	кибернетика_NOUN	0,825
		биофизик_NOUN	0,8022
		биофизика_NOUN	0,7321
		кибернетика_PROPN	0,7212
		генетика_NOUN	0,7202
		физик_NOUN	0,7164
		кибернетикий_NOUN	0,7035
		физика_NOUN	0,7031
		биолог_NOUN	0,6986
		генетик_NOUN	0,6981

Таблица 6

Примеры ошибок, связанных с омонимией корней, при определении наиболее семантически близких слов

Корпус, способ получения векторов и модель морфемного анализа	Слово	10 слов, наиболее семантически близких к данному		Комментарий
Корпус: Araneum **Способ:** *word2vec + частотные веса* **Модель:** свёрточные нейронные сети	восемьсотпятидесятилетие_NOUN	лет_NOUN	0,9735	Корень *-лет-* имеет омонимы, встречающиеся в словах *восемьсотпятидесятилетие, Лета, летать*
		кануть_VERB	0,5956	
		быстрокрылый_ADJ	0,504	
		улетать_VERB	0,4877	
		молниеносно_ADV	0,4777	
		влет_NOUN	0,4686	
		уплывать_VERB	0,4623	
		сизокрылый_ADJ	0,4603	
		лететь_VERB	0,4593	
		влет_ADV	0,4574	
Корпус: Araneum **Способ:** *word2vec + частотные веса* **Модель:** Morfessor	шумозаградительный_ADJ	ленинград_PROPN	0,7059	Корень *-град-* имеет омонимы, встречающиеся в словах *заграждать* и *град* (*город*)
		шум_NOUN	0,6658	
		пенополиуританин_NOUN	0,6201	
		москва::ленинград_PROPN	0,5931	
		тихвин_PROPN	0,5928	
		валяносапожник_NOUN	0,589	
		сталинград_PROPN	0,5872	
		ленинград::москва_PROPN	0,5847	
		псков_PROPN	0,5755	
		петроград_PROPN	0,5726	

ЗАКЛЮЧЕНИЕ

Создание распределённых представлений для редких слов на основе однокоренных слов доказывает свою эффективность для определения семантической близости пар слов. При этом в рамках предложенного подхода необходимо использовать модели морфемного анализа, способные определять тип морфов, так как они сообщают больше информации о морфемном составе слова. Для моделей, которые не имеют такого функционала, предлагается эвристика для определения корня. Полученные представления слов могут применяться для различных задач классификации текстов, в которых встречается много редких слов. Перспективами исследования являются разрешение омонимии корней слов при построении векторов и оценивание качества распределённых представлений на других внешних задачах помимо определения семантической близости пар слов.

СПИСОК ЛИТЕРАТУРЫ

1. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // ICLR: Proceedings of the International Conference on Learning Representations Workshop Track, Arizona. – 2013. – URL: arXiv:1301.3781 [cs.CL] (дата обращения: 20.06.2020).
2. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information // Transactions of the Association of Computational Linguistics. – 2017. – Vol. 5. – P. 135-146.
3. Kutuzov A., Kuzmenko E. WebVectors: A toolkit for building web interfaces for vector semantic models // Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661. – Cham, Switzerland: Springer, 2017. – P. 155-161.
4. Тихонов А.Н. Морфемно-орфографический словарь русского языка. – М.: АСТ, 2002. – 704 с.
5. Smit P., Virpioja S., Grönroos S.A., Kurimo M. Morfessor 2.0: Toolkit for statistical morphological segmentation // The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). – Gothenburg: Aalto University, 2014. – P. 21-24.
6. Virpioja S., Smit P., Grönroos S.A., Kurimo M. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline // Aalto University publication series SCIENCE + TECHNOLOGY. – Helsinki: School of Electrical Engineering, 2013. – Vol. 25. – 32 p.
7. Ruokolainen T., Kohonen O., Virpioja S., Kurimo M. Supervised morphological segmentation in a low-resource learning setting using condi-

tional random fields // Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL). – Sofia: Association for Computational Linguistics, 2013. – P. 29-37
8. Ruokolainen T., Kohonen O., Virpioja S., Kurimo M.. Painless semi-supervised morphological segmentation using conditional random fields // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. – Gothenburg: Association for Computational Linguistics, 2014. – Vol. 2. – P. 84-89.
9. Arefyev N.V., Gratsianova T.Y., Popov K.P. Morphological segmentation with sequence to sequence neural network // Computational linguistics and intellectual technologies: proceedings of the international conference “Dialogue 2018”. – Moscow: Russian State University for the Humanities, 2018. – P. 85-95.
10. Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of russian language // Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science. – Cham: Springer, 2018. – Vol. 930. – P. 3-10.
11. Bolshakova E.I., Sapin A.S. Comparing models of morpheme analysis for russian words based on machine learning // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. – Moscow: Russian State University for the Humanities, 2019. – P. 104-113.
12. Bolshakova E., Sapin A. Bi-LSTM Model for Morpheme Segmentation of Russian Words // Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science. Vol. 1119. – Cham: Springer, 2019. – P. 151-160.
13. Sadov M.A., Kutuzov A.B. Use of Morphology in Distributional Word Embedding Models: Russian Language Case // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” – Moscow: Russian State University for the Humanities, 2018. – URL: http://www.dialog-21.ru/media/4554/sadovmapluskutuzovab.pdf (дата обращения: 20.06.2020).
14. Romanov V., Khusainova A. Evaluation of morphological embeddings for English and Russian languages // Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP. – Minneapolis: Association for Computational Linguistics, 2019. – P. 77–81.
15. Galinsky R., Kovalenko T., Yakovleva J., Filchenkov A. Morpheme level word embedding // Artificial Intelligence and Natural Language. AINL 2017. – Communications in Computer and Information Science. – Vol. 789. – Cham: Springer, 2018. – P. 143-155.
16. Üstün A., Kurfalı M., Can B. Characters or Morphemes: How to Represent Words? – Melbourne: Association for Computational Linguistics, 2018. – P. 144-153.
17. Русская грамматика. – Т. 1: Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология / гл. ред. Н.Ю. Шведова. – М.: Наука, 1980. – 789 с.
18. Maltina L., Malafeev A. morpheme segmentation for russian: evaluation of convolutional neural network models // Analysis of Images, Social Networks and Texts. 8th International Conference, AIST 2019, Revised Selected Papers. Communications in Computer and Information Science.Vol. 1086. – Cham: Springer, 2020. — P. 160-166.
19. Speer R., Chin J., Lin A., Jewett S., Nathan L. Wordfreq: LuminosoInsight/wordfreq: v2.2. Zenodo – 2018. – URL: https://doi.org/ 10.5281/ zenodo.1443582 (дата обращения: 20.06.2020).
20. Panchenko A., Loukachevitch N.V., Ustalov D., Paperno D., Meyer C.M., Konstantinova N. RUSSE: The FIRST WORKSHOP ON RUSSIAN SEMANTIC SIMILARITY // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. – Moscow: Russian State University for the Humanities, 2015. – P. 89–105.
21. Benko V., Zakharov V.P. Very large russian corpora: new opportunities and new challenges // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. – Moscow: Russian State University for the Humanities, 2016. – P. 79-93.
22. Shavrina T., Shapovalova O. To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser // Proceedings of “CORPORA2017”, international conference. – Saint-Petersburg: St Petersburg University, 2017.

Материал поступил в редакцию 15.09.20.

Сведения об авторах

МАЛАФЕЕВ АЛЕКСЕЙ Юрьевич – кандидат филологических наук, доцент, Национальный исследовательский университет "Высшая школа экономики", Департамент прикладной лингвистики и иностранных языков, г. Нижний Новогород
e-mail: amalafeev@yandex.ru

МАЛЬТИНА Людмила Павловна – преподаватель, Национальный исследовательский университет "Высшая школа экономики", Департамент прикладной лингвистики и иностранных языков, г. Нижний Новогород
e-mail: lpmaltina@gmail.com

УДК 004.934.2:164.034:159.946.3

Н.И. Сидняев, Ю.И. Бутенко, Ю.В. Строганов, А.Д. Киселёва

Предикативная симптоматика и биометрия речевого поведения

Рассматриваются вопросы предикативной аналитики при использовании различных методов и алгоритмов для прогнозирования речи на основе статистических данных, а также распознавания речевой информации с помощью нейросетевых обучающихся систем. Приведены вероятностные элементы текста и речевого поведения, которые должны быть учтены при создании алгоритма распознавания речи и выдачи рекомендаций по её улучшению. Для преобразования речевого акустического сигнала в цепочку символов и слов предлагается алгоритм анализа предложений. Анализируются принципы работы современных систем распознавания речи. Предложена методика обработки речевых фраз с использованием математических алгоритмов с оценкой уровней сигналов, позволяющей определить степень влияния индивидуальных особенностей речевого аппарата. Проведена оценка выраженности аномальной симптоматики у испытуемых с дефектами речи и предложен метод количественной оценки выраженности отклонения симптоматики у этих испытуемых.

Ключевые слова: *речевое поведение, биометрия, методика, распознавание речи, симптоматика, эксперимент, статистика, анализ сигнала*

DOI: 10.36535/0548-0027-2021-02-3

ВВЕДЕНИЕ

Система распознавания голоса по уникальным физическим и поведенческим характеристикам (например, слепок голоса) сегодня весьма актуальна и носит название «предикативная аналитика». Согласно современным представлениям структура поведения человека существенно зависит от опоры организма на накопленный и сохраняемый им опыт адаптации к изменчивым условиям среды.

Проблема «прошлого опыта», его структуры и характера использования индивидом занимает наиважнейшее место в современных теориях поведения [1], в которых воздействие среды на организм рассматривается как стохастический процесс [2-4]. Адаптация организма к среде свидетельствует о том, что закономерности среды адекватно отражены в структуре прошлого опыта индивида и используются им при формировании стратегии поведения. Очевидно, что при неадекватном отражении закономерностей среды в прошлом опыте или при некорректном их использовании, поведение индивида перестает быть адаптивным. При экспериментальном изучении особенностей отражения закономерностей среды в прошлом опыте индивида предметом исследования является не «прошлый опыт» вообще, а какой-либо определенный его аспект, т. е. тот или иной комплекс сведений, использование которого позволяет индивиду адекватно вести себя в данной конкретной экспериментальной ситуации [5-7].

В частности, плодотворным оказывается исследование использования индивидом имеющихся у него сведений о вероятностях тех или иных событий во внешней среде, для чего моделируется экспериментальная ситуация, где адекватность поведения индивида конкретизируется параметрами вероятностных характеристик стимульных воздействий. Подобные задачи предикативной аналитики рассматриваются, например, в [8].

Основным методологическим приемом экспериментального исследования различных механизмов поведения является сопоставление поведения в норме и при состояниях, отличающихся от нормы. Особый вклад в современные представления о механизмах речи был сделан путем обобщения наблюдений речевых фрагментов, когда возникают различные виды нарушений, с использованием предикативной аналитики. Для исследования механизмов использования индивидом сведений о вероятностной структуре среды принципиальный интерес представляет анализ тех искаженных состояний, при которых эти механизмы «выведены из строя», т. е. имеет место или изменение вероятностной структуры прошлого опыта, или нарушения механизмов опоры на этот опыт.

При исследовании речевого поведения человека значительное место отводится изучению вероятностных характеристик элементов речи, влияние которых прослеживается в самых разных процессах функционирования речи. Роль вероятностных факторов проявляется в том, что в экспериментах по изучению речевого поведения человека реакция участвующих в исследовании на частые стимулы (слова, слоги, буквосочетания и т.п.) отличается от их реакции на редкие стимулы. Эти наблюдения позволили выдвинуть

гипотезу о том, что в речевых механизмах существует определенная иерархическая организация элементов речи в соответствии с частотой их встречаемости в речевой деятельности. В силу такой организации вероятностные характеристики элементов речи с большей определенностью прогнозируют результаты самых разных операций, связанных с переработкой речевой информации. Каждое слово имеет «индекс частоты», соответствующий частоте встречаемости данного слова в опыте. Способ хранения слова в памяти представляет собой стимул [2, 9].

В качестве конкретной методики для сравнительного изучения вероятностной организации речевого поведения в норме и при нарушении предложен метод субъективных оценок частот слов, который предполагает обращение к прошлому речевому опыту.

Существенным следствием нашей гипотезы является то обстоятельство, что при наличии достаточно надежных оценок частот стимулов можно, в силу общности организации слов по частоте у разных индивидов, прогнозировать результаты экспериментов, в которых характер реакции наблюдаемых на частые стимулы отличается от характера их реакций на редкие стимулы.

ПРЕДИКАТИВНАЯ АНАЛИТИКА И ПОНЯТИЕ «НОРМА». СИСТЕМА РАЗМЕТКИ ЗВУЧАЩЕЙ РЕЧИ

Многочисленные экспериментальные данные, накопленные к настоящему времени, свидетельствуют о том, что в речевых механизмах имеется вероятностная иерархия элементов текста. В самом общем виде это проявляется в том, что наблюдаемые по разному реагируют на частые и редкие вербальные стимулы в ряде экспериментальных условий: при распознавании элементов текста зрительно и на слух и т.д. Отсюда следует, что при изучении того или иного аспекта речевого поведения, независимо от узости задачи исследования, необходимо принимать во внимание вероятностные характеристики тех элементов речи, которыми оперируют в эксперименте.

Так, большинство независимых систем распознавания речи содержат набор акустических моделей (например, скрытые Марковские модели) [3], параметры которых оцениваются при помощи речевых данных большого количества наблюдаемых. В настоящее время выделяют два принципиальных различия между испытуемыми (например, дикторами): акустические, связанные с размером и формой голосового тракта, и различия в произношении, которые обычно называются акцентом. Однако на практике довольно трудно получить полный охват всех акцентов.

В связи с этим, например, в рамках изучения диалектной идентификации на основе распознавания языков с целью идентификации диалектов предполагается использование системы разметки звучащей речи и более глубокого анализа эффективности известных фонотактических подходов. Для этого нами предложено программное обеспечение, а именно – веб-приложение, позволяющее пользователю с помощью браузера взаимодействовать с речевой базой данных, размещенной на удаленном сервере. Интерфейс представляет виджет, создающий визуализацию загруженной аудиозаписи, и набор форм для описания фонем и наблюдаемого. Визуализация аудиозаписи является осциллограммой, которую пользователь может масштабировать и перемещать с помощью полосы прокрутки. Имеется также возможность выделять на осциллограмме необходимые участки для цифровой обработки речевого сигнала в задаче распознавания изолированных слов с применением сигнальных процессов, соответствующих определенным фонемам [9]. Затем в специальную форму заносится информация о времени начала и завершения этого процесса. Далее в этой форме наблюдателю необходимо указать язык, диалект и транскрипцию для рассматриваемой фонемы. Чтобы загрузить размеченные фонемы в базу данных, также необходимо заполнить форму с информацией о испытуемом.

В настоящих исследованиях звуки были разделены на I и II фазы, при этом вторые фазы согласных равны друг другу, и разница между ними заключается только в первой фазе. Причина изменения акцентов – временное удлинение второй фазы, в то время как первая фаза во время разговора остается прежней. Поэтому достаточно наблюдать за звучанием первой фазы распознавания речи, так как продление времени второй фазы не может изменить фонемный состав слова или предложения[10].

Основные сущности или узлы в используемых речевых закономерностях компилировались в соответствующие базы данных – это 'фонема', 'наблюдаемый' и 'нарушение речи'; второстепенные сущности – 'запись', 'страна' и 'город'. Каждая из сущностей хранит определенную информацию, касающуюся фонем и испытуемого: 'фонема' содержит информацию о языке и диалекте, а также транскрипцию; 'диктор' включает полное имя и информацию о поле, возрасте и родном языке; 'нарушение речи' описывает наличие акцента или дефектов речи, если таковые имеются. Подобное разбиение позволяет получать информацию о произношениях фонем по различным заданным параметрам.

Описанная в [10–12] система для разметки звучащей речи является основой для разработки речевого корпуса как совокупности речевых фрагментов и созданных на их основе баз данных, обеспеченных программными средствами доступа к ним. Организован также сбор голосовых записей двух форм: произнесенные в виде отдельных звуков, а также в виде коротких текстов с частым употреблением исследуемых звуков в разных позициях – в начале и конце слова или слога и изучение влияния соседних фонем. Предметом дальнейшего исследования является определение минимальных единиц для сравнения (аллофон, фонема, сочетание нескольких фонем) диалектов, которые необходимо размечать в речевом корпусе. Более того, задача сбора аудиозаписей усложнена рядом таких факторов, как уровень образования наблюдаемого, длительность проживания в другом регионе с ярко выраженным диалектом, наличие физиологических особенностей, социально-культурная общность носителя языка, развитие речевых нарушений, связанных с дыханием (рис. 1).

Рис. 1. Виды речевых нарушений.

Трудность определения понятия «норма» состоит в том, что сама задача оценки субъективных частот слов методом последовательных интервалов не предполагает единственно правильного решения, по отношению к которому все другие решения можно было бы рассматривать как неправильные, не вполне правильные и т. п. Понятие «норма», тем самым, не может быть сформулировано в терминах «правильного» решения [11].

МЕТОД ОБРАБОТКИ СТАТИСТИЧЕСКИХ РЕЗУЛЬТАТОВ В УСЛОВИЯХ ПОМЕХ

Рассмотрим некоторые теоретические проблемы, связанные со спецификой измерения субъективных ощущений. Допустим, что испытуемый №1 поместил слово *челнок* в категорию 2 семиразрядной шкалы употребительности, т. е. шкалы равно кажущихся интервалов, поскольку предполагается, что на психологическом континиуме она делится на равные интервалы, а слово *музыка* – в категорию 7; испытуемый №2 поместил слово *челнок* в категорию 3, а слово *музыка* – в категорию 5. Какой из испытуемых дал лучшее решение? На основе приведенного примера на этот вопрос нельзя дать однозначный ответ, потому что невозможно указать, какие оценки для слов *музыка* и *челнок* являются лучшими. Представим теперь, что из 1000 испытуемых той же социально-культурной группы, что и испытуемые №1 и №2, 995 человек дали те же оценки, что испытуемый №1, и только 5 – оценили слова *челнок* и *музыка*, как испытуемый №2. Нам по-прежнему неизвестно, какие оценки являются лучшими, но мы знаем, что испытуемый №1 действовал так же, как и абсолютное большинство носителей изучаемого языка, а испытуемый №2 – как абсолютное меньшинство, т. е. поведение испытуемого №1 в этом смысле «типично» для поведения здоровых носителей языка, а поведение испытуемого №2 – «нетипично». Таким образом, необходимо сформулировать понятие «норма» в терминах «типичности» решения той или иной конкретной задачи, т. е. соотнести понятие «норма» с некоторой обобщенной картиной поведения группы лиц с правильной дикцией при выполнении определенного задания, а поведение индивида описывать с точки зрения его сходства или отличия от «нормы». Теперь нам следует найти какой-либо способ, с помощью которого можно описать, во-первых, поведение группы в целом, и, во-вторых, поведение индивида сравнительно с поведением группы [12].

В работе [4] отмечено, что многочисленные экспериментальные факты свидетельствуют о том, что речевой опыт человека вероятностно упорядочен. Таким образом, в рамках некоторой социально-культурной общности индивидов – носителей языка – существующая в их речевых механизмах иерархия слов по вероятности является одинаковой [13], т. е. каждое слово имеет в речевой практике этой общности носителей языка вполне определенную вероятность появления X_i. Условимся называть вероятность X_i «истинной вероятностью» слова. Как соотносится с X_i, вероятность x_{ij} слова i, зафиксированная в речевом опыте каждого индивида jиз той же совокупности? Естественно представить, что в общем случае x_{ij} может не совпадать с величиной X_i в точности, а так или иначе отклоняться от неё. Нам представляет-

ся логичным интерпретировать «истинную вероятность» слова X_i как центральную тенденцию того распределения вероятностей $x_{ij_1},\ldots,x_{ij_2},\ldots,x_{ij_n}$., которые существуют в сознании отдельных индивидов – $j_1,\ldots,j_2,\ldots j_n$.

Выборочное исследование при опросе группы индивидов дало для слова *i*индивидуальные оценки вероятностей $x_{ij_1},\ldots,x_{ij_n}$. (переобозначим их $(\overline{x}_{ij_1},\ldots,\overline{x}_{ij_n})$), которые можно рассматривать как реализацию случайной величины X_i. Наилучшую оценку «истинной вероятности» X_i можно получить, используя какую-либо из мер центральной тенденции распределения индивидуальных оценок, например, моду распределения $\overline{x}_{ij}$. Очевидно, что те индивидуальные оценки слова *i*, которые в точности совпадают с модой распределения, могут рассматриваться как наилучшие приближения к «истинной вероятности» X_i. Эти оценки естественно считать «нормой» для слова *i*. Остальные индивидуальные оценки того же слова будут рассматриваться в терминах их отклонений от моды, т. е. в терминах отклонения поведения испытуемых *j* от поведения, характерного для рассматриваемой группы индивидов [12].

Итак, индивидуальная оценка $\overline{x}_{ij}$ тем ближе к «норме», чем большее число испытуемых дают слову *i* ту же оценку, что и индивид*j*. Отличие оценки $\overline{x}_{ij}$ от «нормы» может быть выражено некоторым числом. Поскольку в нашей методике испытуемым предъявляется набор слов, то сравнение индивидуальных оценок с «нормой» проводится для каждого слова из набора, в результате чего получаем ряд чисел, описывающих отличие поведения испытуемого от «нормы» в целом по набору. Принимается, что лучшим приближением к индексам частоты являются субъективные оценки частот слов $F_{суб}$, получаемые в результате психометрического эксперимента с группой индивидов – носителей языка. Таким образом, во многих случаях исследователю важнее знать именно $F_{суб}$ интересующих его слов, а не сведения о частоте этих слов по словарю. Такое определение понятия «норма» приводит нас к следующей процедуре описания поведения отдельного испытуемого сравнительно с поведением группы испытуемых в эксперименте по получению $F_{суб}$ слов [12]:

A. Для некоторого набора слов определяются $F_{суб}$ слов по показаниям большой группы индивидов с правильной дикцией. Полученные распределения оценок рассматриваются как основа для определения «нормы».

B. Степень отличия оценок каждого отдельного индивида от «нормы», т. е. от моды распределения оценок каждого слова набора, описывается с помощью некоторой численной характеристики.

C. Значения введенной численной характеристики вычисляются для каждого испытуемого – члена обследованной группы в целом по набору; по этим данным строится распределение значений характеристики для рассматриваемой группы. Полученное распределение значений рассматривается как выборочное распределение по отношению к генеральной совокупности «нормальные индивиды».

D. Вычисляется оценка вероятности появления любого заданного значения характеристики в выборочной группе «нормальных индивидов» и устанавливается нижний предел этой оценки.

Индивид объявляется не отличающимся по своему поведению от «нормы», если характеристика, описывающая степень его отличия от группы, не выходит за пределы установленного интервала [8]. Наиболее сложным является выбор численной характеристики, удобной для описания степени отличия оценок индивида от оценок группы испытуемых. Поэтому мы начнем рассмотрение процедуры определения нормы именно с этого.

В качестве примера рассмотрим условный набор из пяти слов *a, b, c, d, e*, для которых получены субъективные оценки частот. Матрица распределения оценок для группы из 100 испытуемых приведена в табл. 1. Здесь, наилучшие оценки истинных вероятностей X_a, X_b, X_c, X_d *и* X_e располагаются соответственно в колонках 7, 6, 6, 7 и 2, поскольку именно эти оценки представляют моды распределения индивидуальных оценок для этих слов. При этом оценка 6 ближе к X_a, чем оценка 5; оценка 3 ближе к X_c, чем оценка 4 и т. д.

Оценки выбранных нами слов, полученные от семи индивидов А, Б, В, Г, Д, Е, Ж, представлены в табл. 2. Задача состоит в том, чтобы установить, насколько отличаются оценки каждого из них от «нормы» в целом по всему набору слов, т. е. от моды распределения оценок каждого слова. Сравнение оценок семи индивидов между собой показало, что их поведение весьма различно. Так, индивид А характеризуется тем, что все его оценки совпадают с наилучшими оценками «истинных вероятностей» слов набора, которые мы установили. Часть оценок испытуемых Б и В также совпадает с наилучшими оценками «истинных вероятностей», а другие оценки достаточно близки к ним.

Поведение испытуемого Г отличается от поведения испытуемых А, Б и В тем, что одно слово набора (*b*) попало в ту категорию, в которую его поместили только двое из группы в 100 человек, т. е. $\overline{x}_{bГ}$ весьма далека от моды. Остальные слова испытуемый Г разместил так же, как испытуемый А. Отклонение $\overline{x}_{bГ}$ от наилучшей оценки X_b может быть следствием случайной ошибки индивида Г при оценке частоты слова *b*. Разумеется, наличие отклонений оценок $\overline{x}_{ij}$ от моды для некоторых *i* еще не означает, что *j*отклоняется от «нормы» в целом по набору. Итак, на интуитивном уровне мы можем считать, что оценки испытуемых А, Б, В и Г в целом близки к «норме».

Рассмотрим теперь поведение остальных трех испытуемых: Д, Е и Ж. Испытуемый Д поместил все слова в одну и ту же категорию. Это может объясняться или тем, что все данные слова ему

кажутся одинаковыми по частоте, или же тем, что он не понял инструкции. Так или иначе, очевидно, что оценка $\overline{X}_{id}$ только случайно может совпадать с модой распределения. Испытуемый Ж поместил все слова в те категории, в которые их не поместил ни один из 100 испытуемых. Наконец, испытуемый Е поместил часть слов в те категории, куда их помещало абсолютное меньшинство испытуемых из 100, а остальные слова оценил так же, как испытуемый Ж.Очевидно, что поведение испытуемых А и Ж представляют предельные случаи: А – полное совпадение $\overline{x}_{ij}$ с наилучшими оценками X_i; Ж – максимальное отличие $\overline{x}_{ij}$ от наилучших оценок вероятности X_i.

Таблица 1

Матрица распределения оценок вероятностей поведения испытуемых

Слово	Распределение оценок						
	1	2	3	4	5	6	7
a	1	0	1	10	12	22	54
b	0	2	4	15	24	30	25
c	0	8	12	3	15	40	22
d	0	12	14	15	14	15	30
e	4	61	20	10	3	2	0

Таблица 2

Оценки распределения слов от семи испытуемых

Слово	Испытуемые						
	А	Б	В	Г	Д	Е	Ж
a	7	7	7	7	4	4	2
b	6	7	7	2	4	3	1
c	6	6	7	6	4	5	1
d	7	6	6	7	4	5	1
e	2	2	3	3	4	5	7

Таблица 3

Простое преобразование матрицы распределения оценок группы испытуемых (100- $\overline{x}_{ij}$)

Слово	Распределение оценок							Штраф	
	1	2	3	4	5	6	7	min	max
a	99	100	99	90	88	78	46	46	100
b	100	98	96	85	76	70	75	70	100
c	100	92	88	97	85	60	78	60	100
d	100	88	86	85	86	85	70	70	100
e	96	39	80	90	97	98	100	39	$100 \sum_{\max} = 500$
								$\sum_{\min} = 285$	

Таблица 4

Штрафные баллы семи испытуемых для каждого из слов условного набора

Слово	Испытуемые						
	А	Б	В	Г	Д	Е	Ж
a	46	46	46	46	90	90	100
b	70	75	75	98	85	96	100
c	60	60	78	60	97	85	100
d	70	85	85	70	85	86	100
e	39	39	39	39	90	97	100
Сумма баллов	$\sum A = 285$	$\sum Б = 305$	$\sum B = 364$	$\sum Г = 313$	$\sum Д = 447$	$\sum E = 454$	$\sum Ж = 500$
Значение коэф. *П*	1,00	1,07	1,28	1,10	1,57	1,60	1,75

Теперь задача заключается в том, чтобы найти числовую характеристику, описывающую отличие оценок $\overline{x}_{ij}$ отдельного испытуемого *j* от моды распределения оценок по всем словам *i*. Очевидно, что предельные значения этой характеристики должны соответствовать случаям А и Ж. В качестве такой характеристики предлагается коэффициент парадоксальности $П$.

Для подсчета коэффициента $П$ вначале составляется так называемая «обращенная» таблица, которая является результатом простого преобразования матрицы распределения оценок $\overline{x}_{ij}$. Преобразование состоит в том, что вместо оценок $\overline{x}_{ij}$ (табл. 1) выписываются величины $100 - \overline{x}_{ij}$ (табл. 3).Таким образом, наилучшей оценке вероятности X_i в строке обращенной таблицы соответствует минимальное число. Если индивидуальная оценка $\overline{x}_{ij}$ совпала с модой, то поведение индивида *j* в отношении слова *i* является оптимальным, и за оценку этого слова испытуемому *j* приписывается минимально возможное для слова *i* число «штрафных» баллов. Заметим, что при достаточно большом числе испытуемых практически не встречается ситуация абсолютного единодушия в оценках, поэтому минимальный «штраф» всегда бывает больше нуля. В нашем примере минимальный «штраф» колеблется от 39 (слово *e)* до 70 (слова *b* и *d)*. Если испытуемый поместил стимул в категории, не совпадающие с модой, то ему присваивается больший штраф: в качестве «штрафа» приписывается число из обращенной таблицы (см. табл. 3), соответствующее той категории, в которую испытуемый поместил это слово. В табл. 4 отражены штрафные баллы испытуемых для каждого из слов условного набора.

В соответствии с принятой процедурой, испытуемый А за все слова получил минимальные «штрафы», а испытуемый Ж – максимальные. Для оценки поведения испытуемых в целом естественно взять сумму приписанных им «штрафов» по словам ($\sum А = 285$, $\sum Б = 305$ и т.д.) и вычислить ее отношение к минимально возможной сумме штрафов $\sum$. Указанное отношение индивида представляет коэффициент парадоксальности $П$:

$$П_j = \frac{\sum i}{\sum \min}.$$

Например, значения коэффициента $П$ для примера приведены в последней строке табл. 4. Сопоставим теперь полученные значения $П$ с качественным описанием поведения испытуемых. Очевидно, что чем более сходны оценки испытуемого с модой, тем ближе значение $П$ к 1. Значение $П$ для испытуемого А в точности равно 1, близки к 1 значения $П$ для испытуемых Б, В, Г. Наибольшая сумма «штрафных» баллов получена испытуемым Ж: его коэффициент $П$ достигает максимально возможного значения: $П = 1,75$.

Таким образом, изменение значений коэффициента $П$ соответствует интуитивным представлениям об особенностях поведения испытуемых в ситуации оценки частот слов заданного набора. Для определения нормы выберем набор слов и вычислим относительно этого набора коэффициент $П$ для большой группы испытуемых. Далее следует определить интервал значений $П$, в который укладывается большинство коэффициентов испытуемых, что и будет рассматриваться как норма.

ЭКСПЕРИМЕНТ ПО ПОЛУЧЕНИЮ СУБЪЕКТИВНЫХ ОЦЕНОК ЧАСТОТ СЛОВ

Для эксперимента отобрали 40 существительных, которые в предыдущих исследованиях большинством испытуемых были оценены как частые слова или как редкие. Первые 20 слов списка обычно с большим согласием относятся к классу частых, а остальные 20 – к классу редких. С учетом того, что этот же набор слов будет предъявляться испытуемым с речевым нарушением, все слова должны удовлетворять требованию эмоциональной нейтральности. В эксперименте, по результатам которого определялась «норма», участвовало 100 человек: 50 – это лица с высшим образованием, 20 – со средним и 30 – студенты. Эксперимент проводился по методике последовательных интервалов. Задача испытуемых заключалась в том, чтобы соотнести каждое слово с одной из семи категорий шкалы употребительности. Для фиксации левого конца шкалы в набор были включены «якорные» слова – шесть квазислов. Для каждого слова указано распределение оценок, медиана (M_e) и мода M_o распределения. Чтобы получить распределение коэффициента $П$, характеризующее «норму», были подсчитаны значение $П$ для каждого из 100 испытуемых.

Результаты наблюдений представлены в табл. 5, где указано число испытуемых, для которых получены значения коэффициента $П$, лежащие в данном интервале, и кумулятивные доли испытуемых, у которых $П$ не превышает верхней границы этого интервала.

Из табл. 5 видно, что 90% испытуемых имеют $П \le 1,250$; 96% – $П \le 1,300$.На основе данных табл. 5 представляется разумным определить интервал $П$, соответствующий норме, в пределах от 1,000 до 1,300. Если некоторый индивид получил $П > 1,300$, то он объявляется отличающимся от нормы. Отметим, что в пределах нормы можно выделить интервал значений $П$, в который укладывается абсолютное большинство испытуемых (94% имеют $П \le 1,250$), и интервал, содержащий весьма небольшую долю таких испытуемых (6% имеют $1,250 < П < 1,300$).

Представим теперь содержательную интерпретацию значений коэффициента $П$. При обсуждении условного примера мы отмечали, что $П_{\min}$ соответствует совпадению оценок отдельного испытуемого $\left(\overline{x}_{ij}\right)$ с модой, т. е. наилучшей оценкой «истинной вероятности» X_i; $П_{\max}$ соответствует максимально возможному несовпадению этих оценок, а промежуточные значения $П$ свидетельствуют о той или иной степени отклонения $\overline{x}_{ij}$ от моды. Возникает вопрос, каков характер этих отклонений?

Таблица 5

Таблица кумулятивных процентов испытуемых

Интервал значений $П$	1,000 – 1,100	1,101 – 1,150	1,151 – 1,200	1,201 – 1,250	1,251 – 1,300	1,301 – 1,350	1,351 – 1,400
Число испытуемых, имеющих $П$ в данном интервале	15	26	27	22	6	2	2
Накопленная частота, %	15	41	68	90	96	98	100

Отклоняясь от моды, испытуемый может давать словам оценки на случайном уровне, но, с другой стороны, отклонения могут носить направленный характер: допустим, занижаются частоты частых слов и завышаются частоты редких, или же оценки редких слов лежат много ближе к моде, нежели оценки частых слов и т. п. Можно ли связать те или иные значения $П$ с определенным характером отклонений?

Рассмотрим вначале, какое значение принимает $П$ в случае, когда испытуемый дает оценки слов на случайном уровне. Очевидно, для того, чтобы уровень оценок был строго случайным, поведение испытуемых следует моделировать путем порождения последовательности случайных чисел, имитирующих оценки испытуемых. Для этого с помощью таблицы случайных чисел было выписано 20 случайных последовательностей цифр от 1 до 7. Каждая последовательность содержала 40 элементов – по числу слов в наборе. Таким образом были получены оценки 20 испытуемых в ситуации, когда предполагается, что все испытуемые работают на случайном уровне. Затем для каждого «испытуемого» было подсчитано значение $П$. Найденные значения лежали в пределах $1,400 < П < 1,530$ со средним значением $П = 1,478$ $(\sigma = 0,041)$. В первом приближении мы полагаем возможным принять, что значения коэффициента $П < 1,400$ и $П > 1,530$ свидетельствуют о том, что испытуемый работал на неслучайном уровне. При этом $П < 1,400$ свидетельствует о приближении оценок испытуемых к наилучшим, а $П > 1,530$ – о том, что испытуемый проявляет тенденцию давать оценки, прямо противоположные «норме», т. е. оценивать частые слова как редкие и т. п. В этом случае поведение испытуемого приближается к максимально парадоксальному: для нашего набора $П_{\max} = 1,701$. Из табл. 5. очевидно, что интервал значений $П$, определенный нами как норма, включает только случаи, когда оценки даются на заведомо неслучайном уровне, и даже самое большое наблюдаемое значение $П$, лежащее вне этого интервала, $П = 1,360$ соответствует оценкам на неслучайном уровне. Таким образом, среди 100 испытуемых не нашлось ни одного, который бы работал на случайном уровне или давал оценки, последовательно отклоняющиеся от моды.

Основная задача нашего экспериментального исследования – выяснить, отличаются ли $F_{суб}$ слов, даваемых испытуемым с речевыми нарушениями, от оценок $F_{суб}$ в норме. На основе полученных результатов мы будем судить о том, нарушены ли в речевом поведении испытуемых процессы вероятностного прогнозирования. Чтобы сравнить поведение испытуемых с правильной речью и поведение испытуемых с дефектами речи, следует получить значения коэффициента $П$ для этих двух групп. Естественно ожидать, что данные контрольной группы с правильной речью окажутся в пределах определенной ранее нормы. Если поведение наблюдаемых с дефектами речи окажется вне рамок нормы, то можно будет заключить, что такое поведение при решении задачи получения $F_{суб}$ слов отличается от поведения наблюдаемых с правильной речью. О степени этого отличия будем судить по распределению значений коэффициента $П$ для обеих групп.

Описанный набор из 40 слов был предъявлен группе из 40 лиц, не участвовавших в исходном опыте. Эксперимент проводился по той же методике, что и опыт по определению нормы. Для каждого испытуемого было подсчитано значение $П$. Гистограмма распределения $П$ для контрольной группы из табл. 4 и 5 показывает, что 38 из 40 испытуемых (95%) имеют $П$ в пределах определенной нормы. Это подтверждает разумность выбранного при исследовании критерия.

ЭКСПЕРИМЕНТ В ГРУППЕ С ДЕФЕКТАМИ РЕЧИ. ПРОБЛЕМА ВАЛИДНОСТИ МЕТОДИКИ

Прежде чем проводить эксперимент по получению оценок $F_{суб}$ в группе с дефектом речи, следует учесть некоторые специфические особенности экспериментального исследования таких испытуемых.

При проведении экспериментов в группе с речевыми нарушениями с особой тщательностью должен быть рассмотрен вопрос о валидности применяемой методики. Так, если предложить оценить частоту слов испытуемым без дефектов речи, то полученные результаты можно интерпретировать как выражение опоры исследуемых на их вероятностно упорядоченный прошлый речевой опыт, причем главная тенденция оценок испытуемых и оказывается для нас «нормой» [14]. Однако, проведя тот же эксперимент в группе с нарушениями речи (см. рис. 1) и получив результаты, отклоняющиеся от «нормы», мы не мог-

ли бы априорно утверждать, что эти результаты отражают недостаточную опору с нарушением речи на их прошлый речевой опыт или нарушение вероятностной структуры самого этого опыта. Дело в том, что в эксперименте при распознавании речи (рис. 2) результаты деятельности испытуемых с дефектом речи могут оказаться обусловленными целым рядом факторов, не связанных с механизмами опоры на прошлый опыт или с особенностями структуры этого опыта. Так, испытуемый с дефектом речи может не усвоить инструкцию и выполнять шкалирование слов механически, помещая в любой разряд шкалы любое оказавшееся передним слово. Тот же испытуемый, усвоив инструкцию (о чем можно судить по началу его действий в эксперименте), в дальнейшем может оказаться не в состоянии действовать в соответствии с этой инструкцией, потому что усилившиеся во время эксперимента помехи «отвлекают» его от последовательного выполнения задания. Для того, чтобы оценить результаты деятельности испытуемого с речевым дефектом в эксперименте как отклонения от нормы в интересующем нас аспекте, необходимо по мере возможности исключить влияние на этих испытуемых перечисленных и подобных факторов во время самого исследования.

Контрольным экспериментом с испытуемыми с дефектами речи выбран следующий: испытуемым предлагался набор из 92 карточек, на каждой из которых в случайном порядке было расположено некоторое число точек – от 15 до 74. Задача испытуемых состояла в том, чтобы распределить карточки по девятиразрядной шкале в соответствии с количеством точек на этой карточке: карточки с большим числом точек следовало помещать в правые разряды шкалы (9,8,7...), а карточки с меньшим числом – в левые (1,3...). При этом в инструкции сообщалось, что решение о том, куда поместить карточку, необходимо принимать, определяя число точек на глаз, а не подсчитывая их. Время работы испытуемых ограничивалось. По этой методике было исследовано 10 испытуемых с дефектами речи. Их отбирали в соответствии с современной классификацией типов конечных состояний. Анализ показал, что испытуемые с дефектом речи в конечном состоянии справляются с поставленной задачей примерно так же, как и лица с нормальной речью (коэффициент корреляции Спирмена для результатов по норме и с отклонением составил 0,98). На основании полученных результатов можно считать, что если в эксперименте по оценке слов поведение испытуемых с дефектом речи будет отличаться от поведения лиц без речевых дефектов, то это следует отнести за счет нарушений у испытуемых с дефектом речи вероятностной организации речевого опыта или нарушений опоры на этот опыт. Это дает возможность перейти к постановке основного эксперимента.

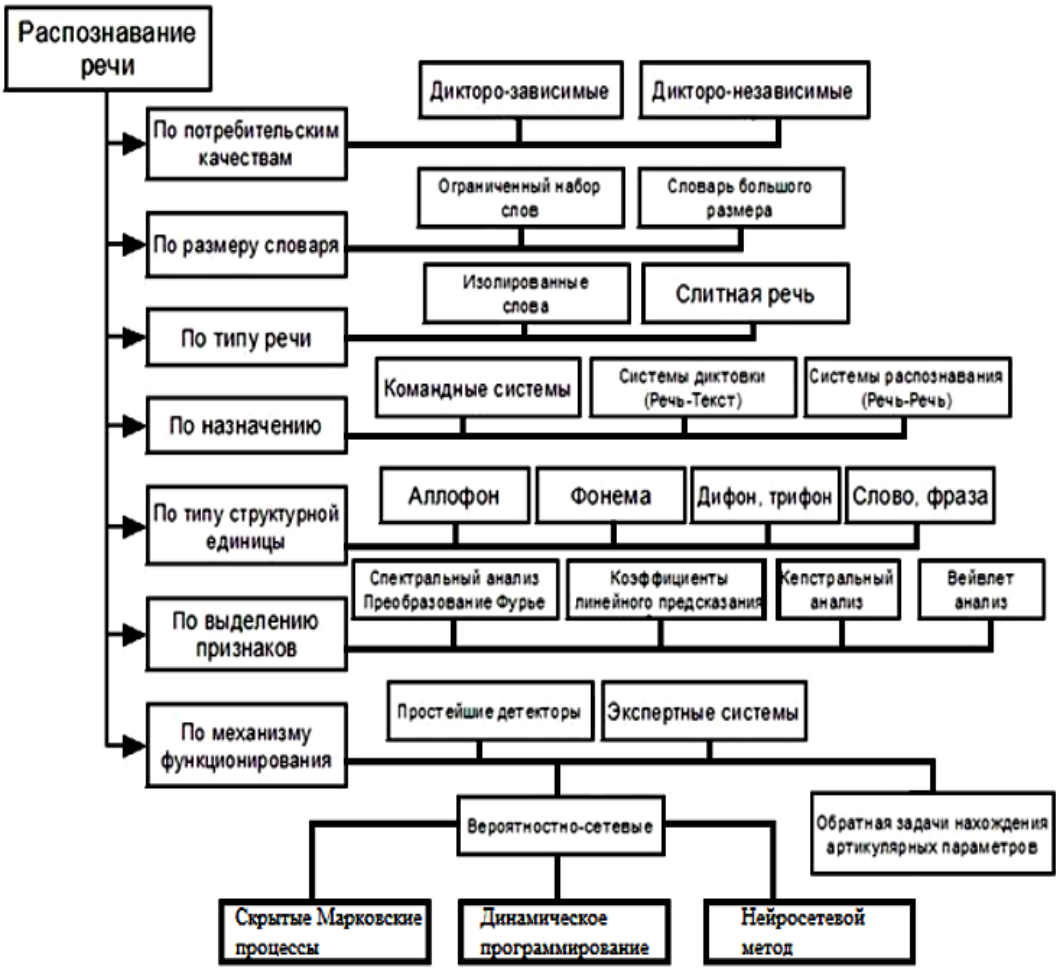

Рис. 2. Методы распознавания речи

ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ ОСНОВНОГО ЭКСПЕРИМЕНТА

Набор из 40 слов был предъявлен группе из 40 испытуемых с дефектом речи. Эксперимент проводился по той же методике, что и в контрольной группе с нормальной речью. Для каждого испытуемого было подсчитано значение коэффициента $П$. Распределения значения $П$ для группы испытуемых отражены в табл. 5. Наблюдаемые значения коэффициента $П$ для испытуемых с дефектом речи располагаются в интервале от 1,149 до 1,526. Подсчет с использованием статистического t-критерия Стьюдента [12] показывает, что среднее значение $П$ в группе с дефектом отличается от среднего значения $П$ в группе с нормальной речью на статистически существенную величину (t=3.19, p<0.01). Очевидно, что оценки в группе испытуемых с дефектом речи отличаются от оценок, полученных в контрольной группе с нормальной речью. Тем не менее, как видно из табл. 4 и 5, интервалы значений $П$ в группе испытуемых с дефектом речи и в контрольной группе с нормальной речью частично перекрывают друг друга: из 40 обследованных испытуемых с дефектом речи у 21 отмечаются значения $П$, лежащие вне пределов нормы, и у 19 – значения $П$ находятся в пределах нормы. Такая ситуация не является неожиданной: согласно литературным данным выборочные распределения оценок для нормы и искажения, получаемые в различных психометрических тестах, в большинстве случаев перекрывают друг друга, т. е. всегда есть некоторая пограничная область значений, куда попадают результаты как лиц с нормальной речью, так и лиц с нарушениями речи.

Разумеется, желательно пользоваться тестом (или экспериментальной методикой), который различает группы с нормой и с нарушением речи без пересечения оценок. Такой тест обладал бы максимальной диагностической силой, но это, скорее, идеал, к которому следует стремиться, а не реальные требования, предъявляемые к большинству экспериментальных методик [15].

Существенно, однако, поставить следующий вопрос: можно ли в рамках рассматриваемой проблемы как-либо интерпретировать тот факт, что значительная часть обследованных с нарушениями речи дает нам те же значения коэффициента $П$, что и лица с нормальной речью? Если согласиться с тем, что выход значения $П$ за пределы нормы свидетельствует о нарушении вероятностной организации речевого опыта, то возможны два варианта интерпретации этого факта:

1) искажение иногда сопровождается нарушением вероятностной организации речевого опыта, а иногда не сопровождается. Такая интерпретация находится в полном соответствии с наблюдаемыми фактами, но не несет ценной информации;

2) дефект речи сопровождается нарушением вероятностной организации прошлого опыта и речевого опыта, в частности. Но степень дезорганизации в среднем пропорциональна степени выраженности психических расстройств: чем больше степень выраженности расстройств, наблюдаемых у испытуемого с нарушением речи, тем с большей уверенностью можно ожидать, что его вероятностный опыт – вообще и речевой опыт, в частности, дезорганизован. Поскольку в нашем исследовании участвовали испытуемые с различными психофизическими состояниями, можно было бы предположить, что группа с дефектами речи, у которой коэффициент $П$ лежит в пределах нормы, характеризуется меньшей тяжестью состояния.

Эта интерпретация содержит объяснение наблюдаемых фактов, но на настоящем этапе исследования должна рассматриваться как гипотеза, требующая экспериментального подтверждения. Для того чтобы подтвердить интерпретацию, следовало бы показать, что парадоксальность речевого поведения, оцениваемая с помощью коэффициента $П$, коррелирована со степенью выраженности эмоционального состояния, наблюдаемого у обследованной группы с дефектом речи.

ОЦЕНКА ВЫРАЖЕННОСТИ АНОМАЛЬНОЙ СИМПТОМАТИКИ У ИСПЫТУЕМЫХ

Общая оценка симптоматики по О–Г шкале выраженности психолингвистических расстройств, позволяющая придать количественные значения степени дефекта, содержит 16 симптомов, характерных для нарушения речи. Эти симптомы были отобраны на основе факторного анализа шкал, прошедших клиническую проверку, т. е. оказавшихся пригодными для оценки эффективности коррекции речи (в баллах). Выраженность каждого симптома подлежит оценке в следующих категориях: «симптом отсутствует» (1 балл); «симптом выражен весьма незначительно» (2 балла); «незначительно» (3 балла); «умеренно» (4 балла); «скорее значительно» (5 баллов); «значительно» (6 баллов); «весьма значительно» (7 баллов). Каждому симптому приписан определенный «вес» на основе специального эксперимента, в котором квалифицированные психологи указывали, насколько у лиц с дефектом речи на разных этапах может быть выражен тот или иной из 16 симптомов. Для определения выраженности каждого симптома её оценка в баллах умножается на число, соответствующее «весу» симптома. Минимальная теоретически возможная общая оценка по О–Г шкале составляет 39 баллов (случай, когда у исследуемого лица отсутствуют все 16 симптомов), а максимальная – 273 балла (случай, когда каждый из симптомов выражен «весьма значительно»); «срединная» оценка составляет160 баллов.

В формулировке симптомов, как правило, психолингвисты не придерживаются традиционных клинических категорий, описывающих симптоматику дефективности речи, поэтому О–Г шкала была подвергнута нами определенной модификации, цель которой – сопоставить пунктам шкалы («симптомам») общепринятые обозначения.

Таким образом, с помощью модифицированной нами О–Г шкалы оказалось возможным придать количественную определенность степени выраженности симптоматики у каждого испытуемого с дефектом речи на период его исследования по методике $P_{(суб)}$.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Ранее указывалось, что из 40 испытуемых с дефектом речи, обследованных по методике с использованием $F_{суб}$, у 21 были отмечены значения *П*, лежащие вне пределов нормы (назовем это группой З), и у 19 – значения *П* находились в пределах нормы (группа С). Необходимо было выяснить, имеют ли группы З и С существенные клинические отличия друг от друга, в частности, отличаются ли эти группы по степени выраженности симптоматики. Сопоставление экспериментальных и клинических данных показало, что в группе З (21 чел.) степень выраженности симптоматики у 20 человек превысила 160 баллов и лишь у одного человека составила 135 баллов. В группе С (19 человек) степень выраженности симптоматики у 17 человек не превысила 158 баллов и лишь у двух – оказалась несколько выше (160 баллов из 163). Коэффициент ранговой корреляции Спирмена между значениями *П* и оценками по шкале О–Г составил 0,74 (значим на 99% уровне).

Таким образом, следует отметить, что между степенью парадоксальности речевого поведения испытуемых и степенью выраженности отклонений правильности их речи имеется отчетливая корреляционная связь. Это дает основание считать, что степень дезорганизации вероятностной структуры речевого опыта в среднем пропорциональна выраженности симптоматики [16]. Тем самым, интерпретация может считаться экспериментально подтвержденной. Поэтому представляется естественным, что испытуемые с дефектом речи со сравнительно небольшой степенью выраженности психопатологической симптоматики могут давать коэффициент, значение которого лежит в пределах эталона нормы.

Отметим, что система распознавания речи состоит из двух моделей: акустической и лингвистической. Компьютер записывает звук речи в виде цифрового сигнала и делит его на отпечатки длительностью несколько миллисекунд. Акустическая модель отвечает за преобразование речевого сигнала в набор признаков, в которых отображена информация о содержании речевого сообщения. Предложенная авторами программа выполняет сложный анализ речи, сравнивая отпечатки с записанными в память речевыми образцами. Лингвистическая модель анализирует информацию, получаемую от акустической модели, и формирует окончательный результат распознавания. На основе вероятностного расчета компьютер определяет, что именно мог произнести испытуемый. В основе этих моделей лежит понятие фонемы – наименьшей акустической единицы языка. Как правило, процесс непрерывного распознавания речи дает до 95% качества распознавания при оптимальных условиях, но, тем не менее, все-таки дает на 100 знаков 4–5 ошибок. Увеличение вычислительных мощностей мобильных устройств позволило создать программы с функцией распознавания речи. Интеллектуальные речевые системы позволяют автоматически синтезировать и распознавать речевой сигнал [18].

В качестве примера проведен эксперимент с группой студентов, среди которых была студентка, её голос заранее записали и получили его отпечаток, сохранив в двоичной форме. Если ставится задача нахождения студентки в группе, то сравнивается ее сохраненный голос с голосом каждого студента из группы. Совпадение голоса с голосом одного студента из этой группы, позволяет сделать вывод, что это и есть интересующая нас студентка. Но, и в этом случае нет уверенности на сто процентов, что это та студентка. Поэтому необходима верификация, т.е. проверка, для этого берется еще раз отпечаток голоса, и проверяется, принадлежит ли этот голос искомой студентке.

Необходимо отметить, что при распознавании речи нельзя забывать о шуме, который ухудшает работу систем распознавания и его невозможно отфильтровать, так как он распространяется по всему сигналу [20].

ЗАКЛЮЧЕНИЕ

В настоящей статье проведены исследования с использованием биометрических показателей голоса, в частности, когда создается такая экспериментальная ситуация, где адекватность поведения исследуемого обусловливается знанием вероятностных характеристик стимульных воздействий. Разработан методологический прием изучения речевого поведения, базирующийся на сопоставлении речи в норме и при состояниях, отличающихся от неё. Для исследования механизмов использования индивидом сведений о вероятностной структуре среды принципиальный интерес представляет анализ тех искаженных состояний, при которых эти механизмы «выведены из строя», т. е. имеет место или изменение вероятностной структуры прошлого опыта, или нарушения механизмов опоры на этот опыт.

В качестве конкретной методики в работе для сравнительного исследования вероятностной организации речевого поведения в норме и при нарушениях был использован метод субъективных оценок частот слов, который предполагает обращение к прошлому речевому опыту. Мы сформулировали само понятие «норма» применительно к некоторой конкретной ситуации, а именно – к поведению правильных речевых фрагментов в наблюдениях по получению субъективных оценок частот слов. В рамках исследования по диалектной идентификации мы использовали системы разметки звучащей речи для более глубокого анализа эффективности известных фонотактических подходов на основе распознавания языков с целью идентификации диалекта.

В нашем исследовании акцент ставился на анализе и моделировании просодической структуры диалектов. Причем просодический подход основан на учении об ударении, изучающем слоги по их ударности и протяженности. Диалекты языка проявляют существенные отличия друг от друга с точки зрения особенностей их просодической структуры, включая различия в их ритмической структуре, темпе речи и длительности гласных звуков. Предложенные просодические отличительные приемы можно использовать со значительной точностью для автоматической идентификации диалекта говорящего, а подход моделирования может значительно улучшить систему распознавания речи, которая использует фонотактические закономерности.

Экспериментальные исследования подтверждают представления о том, что человек владеет вероятностными закономерностями речи и использует имеющиеся у него сведения для оптимизации стратегии речевого поведения. Это означает, что в рамках некоторой социально-культурной общности индивидов – носителей языка – существующая в их речевых механизмах иерархия слов по вероятности одинакова. Так как в предлагаемой нами методике испытуемым предъявляется набор слов, то сравнение индивидуальных оценок с «нормой» проводится для каждого слова из набора, в результате чего получаем ряд чисел, описывающих отличие поведения испытуемого от «нормы» в целом по набору.

Таким образом, в настоящей статье:

- в качестве примера приведен условный набор из нескольких слов, для которых получены субъективные оценки частот;
- установлено, насколько отличаются оценки каждого испытуемого от «нормы» в целом по всему набору слов, т. е. от моды распределения оценок каждого слова. Сравнение оценок индивидов между собой показало, что их поведение весьма различно;
- найдены числовые характеристики, описывающие отличие оценок отдельного наблюдаемого от моды распределения оценок;
- одной из характеристик предложен коэффициент парадоксальности, изменение значений которого соответствуют интуитивным представлениям об особенностях поведения испытуемых в ситуации оценки частот слов заданного набора.

Необходимо отметить, что между степенью парадоксальности речевого поведения испытуемых и степенью выраженности отклонений от правильной их речи имеется отчетливая корреляционная связь. Это дает основания считать, что степень дезорганизации вероятностной структуры речевого опыта в среднем пропорциональна выраженности симптоматики. Поэтому представляется естественным, что испытуемые с дефектом речи со сравнительно небольшой степенью выраженности психопатологической симптоматики могут давать коэффициент, значение которого лежит в пределах нормы.

СПИСОК ЛИТЕРАТУРЫ

1. Рабинер Р.Л., Шафер Р.В. Цифровая обработка речевых сигналов / пер. с англ. под ред. М.В. Назарова, Ю.Н. Прохорова. – М.: Радио и связь, 1981. – 496 с.
2. Филичева Т.Б., Чевелева Н.А., Чиркина Г.В. Основы логопедии : учеб. пособие. – М.: Просвещение, 1989. – 223 с.
3. Фролов А.В., Фролов Г.В. Синтез и распознавание речи. Современные решения. – М.: Связь, 2003. – 216 с.
4. Галунов В.И. Современные проблемы в области распознавания речи. – URL: http://auditech.ru/page/darkness.html (дата обращения: 12.03.2016).
5. Карпов А.А., Кайа Х., Салах А.А. Актуальные задачи и достижения систем паралингвистического анализа речи // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – Т. 16, № 4. – С. 581-592. DOI:10.17586/2226-1494-2016-16-4-581-592
6. Горшков Ю.Г., Дорофеев А.В. Речевые детекторы лжи коммерческого применения // «ИНФОРМОСТ» – «Радиоэлектроника и Телекоммуникации». – 2003. – № 6(30). – С. 13-15.
7. Montacie C., Caraty M.-J. Prosodic cues and answer type detection for the deception subchallenge // Proceedings Interspeech-2016. – San Francisco: STIH Laboratory, Paris Sorbonne University, 2016. – P. 2016-2020. DOI: 10.21437/Interspeech.2016-33.
8. Басов О.О., Карпов А.А., Саитов И.А. Методологические основы синтеза полимодальных инфокоммуникационных систем государственного управления. – Орел: Академия ФСО РФ, 2015. – 271 с.
9. Бутенко Ю.И., Строганов Ю.В., Шевченко В.И., Славнов Н.В., КвасниковА.В. Система разметки звучащей речи для сравнительного анализа произношения в различных диалектах // Вестник Воронежского государственного университета. Серия: Системный анализ иинформационные технологии. – 2020. – №4. – С.168-176. DOI: https://doi.org/10.17308/sait.2020.1/2631
10. Slavnov N.V., Stroganov Y.V., Kvasnikov A.V. System for Speech Corpus Development // IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2020. – P. 1730-1732.
11. Евсеев А.И., Нгуен В.Х. Исследование и разработка метода цифровой обработки речевого сигнала для получения динамики форматных характеристик звуков слова // Вестник московского энергетического института. – 2010. – №4. – С.45-49.
12. Сидняев Н.И., Бутенко Ю.И., Гаража В.В. Статистическая оценка ассоциативной силы неосмысленных буквосочетаний // Теоретическая и прикладная лингвистика. – 2019. – №5(4). – С.107-124.
13. Грачев А.М. Лингвистические подходы к автоматическому распознаванию речи // Вестник нижегородского университета им. Н.И. Лобачевского». – 2013. – № 6-2. – С. 61-63.
14. Магдиева З.Х., Орлов И.М., Беленко М.В. Особенности распознавания дефектной речи современными системами автоматического распознавания речи // Научно-технический вестник Поволжья. – 2019. – №6. – С. 85-87.
15. Санников В.Г. Статистический анализ цифрового синтеза речевого сигнала на основе его авторегрессионной модели // DSPA: Вопросы применения цифровой обработки сигналов. – 2018. – Т.8, №2. – С. 171-176.

16. Петрова Е.М., Ничушкина Т.Н. Анализ проблем автоматического распознавания русской речи // Технологии инженерных и информационных систем. –2019. – №1. – С. 3-10.
17. Кухтинова М.С., Позолотина Н.А., Трубин В.Г. Системы распознавания речи // Автоматика и программная инженерия. – 2014. – № 2(8). – С. 47-49.
18. Тампель И.Б. Автоматическое распознавание речи – основные этапы за 50 лет // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Т.15, №6. – С. 957-968.
19. Будков В.Ю., Савельев А.И., Вольф Д.А. Методика исследования параметров речевого сигнала, отражающая истинность передаваемой информации // Доклады ТУСУР. – 2016. – Т. 19, № 2. – С. 56-60. DOI:10.21293/1818-0442-2016-19-2-56-60
20. Бабаринов С.Л., Будникова М.А. О распознавании речи // Научные ведомости Белгородского государственного университета. Серия: «Экономика. Информатика». – 2014. – № 21(192). – С. 182-185.

Материал поступил в редакцию 11.09.20.

Сведения об авторах

СИДНЯЕВ Николай Иванович – доктор технических наук, профессор, заведующий кафедрой Московского государственного технического университета имени Н.Э. Баумана (Национальный исследовательский университет)
e-mail: Sidn_ni@mail.ru

БУТЕНКО Юлия Ивановна – кандидат технических наук, доцент Московского государственного технического университета имени Н.Э. Баумана (Национальный исследовательский университет).
e-mail: iuliiabutenko2015@yandex.ru

СТРОГАНОВ Юрий Владимирович – старший преподаватель Московского государственного технического университета имени Н.Э. Баумана (Национальный исследовательский университет).
e-mail:stroganovyv@bmstu.ru

КИСЕЛЁВА Аполлинария Дмитриевна – студент Московского государственного технического университета имени Н.Э. Баумана (Национальный исследовательский университет).
e-mail: apollinariakis@mail.ru

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 81’322.4’373.46

В.И. Хайруллин

Проблема переводческих универсалий (от уровня терминологии до уровня переводческих трансформаций)

Переводческие универсалии рассматриваются как закономерности, общие для всех случаев перевода с одного языка на другой. Высказывается предположение о возможности обнаружения таких универсалий на уровнях более высоких, чем уровень терминологии. Продуктивным представляется исследование уровня переводческих трансформаций как приема, с помощью которого осуществляется переход от единиц оригинала к единицам перевода. Подчеркивается, что понятие и феномен перевода носят характер универсальности.

Ключевые слова: *термин, терминология, уровень, прием, трансформация, расширение, перевод, исходный язык, язык перевода*

DOI: 10.36535/0548-0027-2021-02-4

Широкая ориентированность и возможность обретения неординарных решений делают информационную сферу одной из наиболее привлекательных для исследования [1–2]. Для нас особенно ценно то, что она позволяет рассмотреть в своих терминах такую актуальную проблему современной науки, как проблема языковых универсалий, корнями уходящую в античные грамматики с их учениями о членах предложения и частях речи. Более поздними, средневековыми, «отцами» понятий универсальности признаются Я.А. Коменский и Р. Бэкон, а также аббаты А. Арно и К. Лансло, разработавшие в 1660 г. «Грамматику Пор-Рояль». Именно этот труд относится к универсальным, в которых категории грамматики трактуются и эксплицируются посредством категорий мышления и восприятия [3, с. 6].

Интерес к универсалиям стал особенно очевиден в XX веке под влиянием работ Н.С. Трубецкого [4], Р.О. Якобсона [5] и в дальнейшем – Дж. Гринберга [6].

Если принять положение о том, что языковыми универсалиями служат закономерности, общие для всех или многих языков, то можно предположить, что понятие универсалии распространимо на все области, связанные с проявлением языковой деятельности, в частности с областью перевода. В таком случае, возникают вопросы: действительно ли существуют переводческие универсалии, а если они существуют, возможно ли дать им определение?

Отвечая на эти вопросы, следует признать, что закономерности наблюдаются на всех уровнях языковой системы. Чем ниже этот уровень, тем ярче эти закономерности обнаруживаются. И наоборот, высокие уровни не дают возможности однозначного выявления закономерностей при рассмотрении их функционирования в разных языках. К таким высоким и сложным уровням относится межъязыковая коммуникация, а именно – перевод. Исследование переводческих универсалий осложняется тем, что анализ требует доступа как к оригинальным текстам, т. е. текстам на исходном языке (ИЯ), так и к текстам на языке перевода (ПЯ). Выявляемые при этом особенности, например интерференция ИЯ в ПЯ, выражающаяся в использовании в ПЯ нетипичных лексических единиц, лексико-грамматических комбинаций и речевых структур, некоторым исследователям может представляться в качестве одной из искомых универсалий. С этим трудно согласиться, поскольку интерференция сигнализирует о несоблюдении одной или нескольких норм ПЯ, в частности, стилистической, узуальной или прагматической. В результате текст воспринимается реципиентом не как аутентичный, а как переводной, причем выполненный с нарушением перечисленных норм. По этой причине интерференция, часто как следствие стилистической неаккуратности переводчика, не может рассматриваться в качестве универсалии. При этом следует отметить, что воспринимающая аудитория может иметь

свои предпочтения в отношении переводного текста – в ряде случаев реципиенты предпочитают читать переводной текст как аутентичный, т. е. следующий нормам языка перевода, тогда как в иных ситуациях реципиенты могут проявлять благосклонность в отношении лексических единиц текста, в частности, терминов, имеющих иноязычное происхождение, и относиться к ним лояльно.

Более плодотворным с точки зрения поиска переводческих универсалий является один из переводческих приемов, получивший название расширения, экспликации, т. е. указания в переводе дополнительных признаков или терминов, которые как бы разъясняют передаваемую информацию. Этот прием часто используется в языке науки и техники. Например, при переводе следующего высказывания с английского языка на русский расширение достигается за счет использования таких терминов, как «нацелено», «технологического», «бурения»:

For each particular task, a solution has been defined that improves all processes – from the well to the refinery.

Для каждой конкретной задачи разработано решение, которое нацелено на совершенствование технологического процесса – от бурения скважины до нефтеперерабатывающего завода.

Сопоставительный анализ показывает, что при переводе научно-технического текста таких дополнительных терминов по отношению к терминам ИЯ насчитывается до 11%. Хотя эта цифра весома, но едва ли она позволяет говорить об универсальном характере приема экспликации. Расширение имеет место, однако не в подавляющем большинстве переводов.

Кроме того, переводческую экспликацию следует рассматривать отдельно от иных случаев расширения, которые имеют место в непереводных, аутентичных текстах, широко использующих иноязычную терминологию, но рассчитанных на широкую аудиторию, требующую более детального описания, причем средствами русского языка. Например, использование расширительного описания «снижение уровня притязаний» наряду с используемым в тексте термином «дауншифтинг» (от английского *downshifting* – буквально «перемещение вниз»). Если подобное расширение имеет почти регулярный характер в аутентичных текстах [7, с. 29], то в переводных, повторяем, лишь в 11% случаев. Использование приема расширения зависит как от функционального стиля текста, так и от целевой аудитории, т. е. от реципиентов перевода.

Существует также точка зрения, в соответствии с которой к категории универсалий относят перевод непереводимого, т.е. способы описания средствами языка перевода реалий, которые наличествуют в исходной культуре, но отсутствуют в целевой. Такие единицы называются уникальными, не находящими себе пары в других языках и поэтому отсутствующими в ментальном глоссарии билингва.

В качестве примера приведем русский термин «сутки» – никогда однозначно не переводимый в большинстве европейских языков. В английском языке термин обычно передается с использованием приема расширения: *twenty four hours*, буквально «двадцать четыре часа». Напротив, английский термин *fortnight* не имеет однозначного соответствия в русском языке, и передается также эксплицитно: *fortnight* – «две недели». Таким образом, попытка найти универсалии на уровне уникальных культурно-языковых единиц приводит нас к описанному приему расширения.

В свое время мы высказывали сомнение в самой возможности переводческих универсалий [8, с. 77]. Однако по прошествии ряда лет такая категоричность представляется малооправданной.

Если задать себе цель найти переводческие универсалии, то, очевидно, их следует искать на более высоких уровнях, не ограничиваясь уровнем терминологии.

На наш взгляд, универсальным является само понятие перевода – именно перевода, а не локализации. Последний термин стал широко использоваться в современных научных изданиях и трактуется как перевод текстов с языка транснациональных корпораций на языки отдельных культур. Последователи такого подхода признают наличие интернационального или глобального исходного текста, транспонируемого в отдельные – локальные языки. При этом перевод понимается как адаптация текста, призванная сделать его приемлемым для понимания носителями локальных языков. Однако следует учитывать, что такая адаптация базируется на переводе, который, собственно, дает основание утверждать, что существуют такие понятия и термины как глобализация и локализация. Иными словами, без перевода не было бы ни первого, ни второго.

Вид языкового посредничества по преобразованию содержания текста на исходном языке в текст на языке перевода, в коммуникативном, информативном и культурном отношениях тождественный тексту на исходном языке, носит название перевода, который служит платформой, поддерживающей модернистские попытки преобразовать перевод в понятия локализации, глобализации или переписывания (*re-writing*) с одного языка на другой: наличие понятия «перевод» порождает многие производные понятия.

При всей безусловной универсальности понятия «перевод» – и в этом заключается его парадоксальность, которая, впрочем, подтверждает его универсальный характер, – он может быть немотивирован, т.е. он может проявлять свою нецелесообразность при рассмотрении некоторых пар языков. Так, немотивированность перевода очевидна при анализе перевода с датского языка на шведский. Эти языки обладают настолько близкородственными отношениями, что перевод с одного на другой оказывается лишенным мотивации: носители датского и шведского языков прекрасно понимают друг друга, не прибегая к переводу, перевод им не нужен, у них нет мотивации в нем. Более того, дабы избежать так называемой датско- или шведоцентричности, т. е. демонстрации своей приверженности к той или другой культуре, участники диалога нередко прибегают к «нейтральному» английскому языку и ведут общение на нем, т. е. в данном случае перевод также оказывается нецелесообразен.

Универсальным представляется и понятие переводческой трансформации, т. е. приема, с помощью которого осуществляется переход от единицы оригинала к единицам перевода [9, с. 248]. Насчитывается целый ряд таких трансформаций. Это генерализация, импликация, смысловое развитие, антонимический

перевод, адвертивный перевод, гендерная трансформация и др., а также упоминавшийся ранее прием экспликации, или расширения. Каждая из названных трансформаций не может претендовать на роль переводческой универсалии, поскольку в отдельности они не находят проявления во всех или в подавляющем большинстве случаев перевода. Это было продемонстрировано на примере переводческого расширения. Вместе с тем, трансформация как переводческая категория, как понятие, в терминах которого возможно описать лексико-семантико-структурные преобразования, наблюдаемые при анализе переводческого материала, соответствует требованиям универсальности – трансформация как таковая наблюдается при сопоставлении переводов с любого языка на любой другой. Более того, понятие трансформации хорошо согласуется с основным постулатом теории перевода, в соответствии с которым данная область исследований должна носить не прескриптивный (предписывающий), а дескриптивный (описательный) характер: мы анализируем то, что мы имеем, то, что получаем в результате перевода, в результате переводческих преобразований, трансформаций. Использовать те или иные из этих тансформаций волен каждый переводчик, свободный от какого-либо диктата обязательств по применению определенных приемов.

Таким образом, переводческие универсалии следует искать на высоких уровнях языковой системы, не ограничиваясь лексико-терминологическим уровнем. К переводческим универсалиям относятся непосредственно понятие и феномен перевода, а также переводческие трансформации, имеющие место во всех и в каждом случае перевода.

СПИСОК ЛИТЕРАТУРЫ

1. Гиляревский Р.С. Информационная сфера: Краткий энциклопедический словарь. – СПб: Профессия, 2016. – 304 с.
2. Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А. О природе информации // Научно-техническая информация. Сер. 2. – 2009. – № 1. – С. 1-6; Belonogov G.G., Gilyarevskii R.S., Khoroshilov A.A. On the nature of information // Automatic Documentation and Mathematical Linguistics. – 2009. – Vol. 43, № 1. – P. 1-6.
3. Успенский Б.А. Проблема универсалий в языкознании // Новое в лингвистике. Вып.5. – М.: Прогресс, 1970. – С. 5-30.
4. Веденина Л.Г. Н.С. Трубецкой в контексте русских лингвокультурных традиций // Филологические науки. – 2015. – № 4. – С. 84-104.
5. Якобсон Р. Типологические исследования и их вклад в сравнительно-историческое языкознание // Новое в лингвистике. Вып. 3. – М.: Прогресс, 1963. – С. 95-105.
6. Гринберг Дж., Осгуд Ч., Дженкинс Дж. Меморандум о языковых универсалиях // Новое в лингвистике. Вып. 5. – М.: Прогресс, 1970. – С. 31-44.
7. Хайруллин В.И. Терминология и локализация: насколько русифицируются терминологические единицы при переводе // Научно-техническая информация. Сер. 2. – 2020. – № 2. – С. 27-30; Khairullin V.I. Terminology and localization: how terminological units are russified during translation // Automatic Documentation and Mathematical Linguistics. – 2020. – Vol. 54, № 1. – P. 52-54.
8. Khairoulline V. Translation universals: do they exist? // Perspectives: Studies in Translatology. – 2008. – Vol. 16, № 1-2. – P. 75-77.
9. Комиссаров В.Н. Теория перевода. – М.: Высшая школа, 1990. – 254 с.

Материал поступил в редакцию 16.12.20.

Сведения об авторе

ХАЙРУЛЛИН Владимир Ихсанович – доктор филологических наук, профессор, профессор кафедры международного права и международных отношений института права Башкирского государственного университета, г. Уфа.
e-mail: vladimir-blt@mail.ru