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. **№ 6** **Москва 2018**

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.65–047.44:[002:530.145]

В.О. Толчеев

Наукометрический анализ современного состояния и перспектив развития квантовых технологий

Проводится наукометрический анализ быстро развивающейся междисциплинарной области, охватывающей вопросы создания и развития квантовых технологий. Дается общая характеристика рассматриваемой тематики и выявляются ключевые направления исследований. Определяются оценки публикационной активности различных стран в период 2000-2016 годов на базе анализа статей, проиндексированных в БД Web of Science. Выделяются страны-лидеры (США и КНР). Рассматриваются наиболее активно прогрессирующие научные направления. Строится прогноз для США и КНР по ожидаемому числу публикаций в краткосрочной перспективе. Особое внимание уделяется публикационной активности российских ученых по тематическим разделам рубрикатора Web of Science. Исследуется интенсивность международного сотрудничества в сфере квантовых технологий и степень интернационализации исследований.

Ключевые слова: *наукометрический анализ, реферативная база данных Web of Science, квантовые технологии, количественные оценки, библиографический документ*

ВВЕДЕНИЕ

В настоящее время в развитых странах активизировались научные исследования в области квантовых технологий (КТ – *Quantum Technologies*). КТ – обобщающий «зонтичный» термин, под которым обычно понимаются устройства, алгоритмы, системы, основанные на использовании квантовых эффектов и предусматривающие манипуляции элементами на квантовом уровне. Исследования ведутся на стыке большого числа дисциплин – физики, химии, информатики (*Computer Science*), математики, оптики, новых материалов, сверхпроводимости, нанотехнологий.

Ряд экспертов ожидают в ближайшее десятилетие «квантовой революции», которая приведет к радикальным изменениям в сфере высокопроизводительных вычислений, передачи и защиты информации, сенсорных и метрологических устройств [1, 2]. Важная особенность КТ заключается в стимулирующем воздействии на большое число «связанных» научных тематик, которые зависят от надежности передачи данных, скорости вычислений, точности измерений, размеров датчиков и сенсоров. К числу областей, которые могут получить наибольший выигрыш от развития КТ, относят вычислительную технику и связь, фотонику, медицину, материаловедение, проектирование и моделирование сложных систем, робототехнику. Уже в настоящее время появились первые коммерчески успешные прикладные разработки на базе КТ (например, дисплеи на квантовых точках – *quantum dots*) [3]. В Китае между Пекином и Шанхаем создана первая в мире линия оптоволоконной квантовой связи протяженностью около двух тысяч километров и развернуты работы по введению в эксплуатацию новых участков сверхзащищенной линии между другими крупными городами. В КНР также запущен первый в мире спутник квантовой связи «Мо-Цзы» и проведены первые эксперименты по квантовому распределению криптографических ключей с помощью «Мо-Цзы» [4].

Канадская фирма *D-Wave Systems* сумела существенно улучшить конструкцию вычислительного устройства, которое, как считает ряд экспертов, может стать прототипом будущего квантового компьютера [5]. Несмотря на значительную критику этой разработки со стороны научного сообщества американские высокотехнологические гиганты *Lockheed Martin и Google* закупили и тестируют компьютеры *D-Wave*.

Не исключено, что в долгосрочной перспективе (не ранее 2030 г.) универсальные квантовые компьютеры смогут выполнить алгоритм Шора, который позволит взломать большинство существующих секретных кодов за счет быстрого разложения больших чисел на простые множители. Одним из первых неэффективным станет асимметричный криптографический алгоритм с открытым ключом — *RSA*. Для купирования возможной опасности Национальный институт стандартов и технологий США в 2016 г. объявил конкурс научных проектов по разработке новых методов зашифровки данных, способных обеспечить защищенность информации в постквантовую эру [6].

Дальнейший прогресс КТ может быть достигнут благодаря запуску крупных специализированных инициатив. Так, в Евросоюзе в дополнение к программам, выполняемым на национальном уровне, решено приступить в 2018 г. к реализации общеевропейского флагманского проекта «*Quantum Technologies Fla*gship», рассчитанного на 10 лет и имеющего бюджет около 1 млрд евро [1]. Аналогичные программы осуществляются в США, Китае, Японии, Республике Корея, Германии, Великобритании, Франции. Наряду с бюджетными исследованиями значительная часть НИОКР финансируется крупнейшими высокотехнологичными компаниями, ориентированными на создание инновационных ноу-хау (прежде всего, *Google, Microsoft, IBM, Apple, Bosch, Siemens, Thales, Toshiba, Hitachi, Samsung, Alibaba*). В развитых странах расширяется число стартапов, организованных для проектирования и коммерциализации КТ.

В России большая часть работ по квантовым технологиям проводится в рамках проекта «Создание технологии обработки информации на основе сверхпроводящих кубитов». Этот проект поддерживается Министерством образования и науки, Фондом перспективных исследований и корпорацией «Росатом» [7, 8]. Ряд передовых исследований выполняется также в Российском квантовом центре, созданном в Сколково (*http://www.rqc.ru*).

К числу ключевых быстро развивающихся направлений в области КТ относят следующие тематики [1, 2, 9]:

1) универсальные квантовые компьютеры (*Quantum Computers*), способные реализовывать различные вычислительные операции, в частности, выполнять алгоритм Шора. На современном этапе основные исследования пока сосредоточены лишь на поиске наиболее подходящей базовой технологий реализации кубитов и обеспечения их устойчивой когерентности;

2) квантовые симуляторы (*Quantum Simulators*), осуществляющие отдельные вычислительные операции, например, операции быстрого поиска и перебора, которые важны для решения, например, оптимизационных задач (именно к данной категории, как представляется, правильно относить разработку канадской фирмы *D-Wave*, которая иногда ошибочно называется «квантовым компьютером»);

3) квантовая передача данных (*Quantum Communication*), получившая на современном этапе значительное прикладное применение. В ведущих странах, прежде всего Китае и США, уже практически сформировался новый сектор высокотехнологичной промышленности, занятый выпуском компонент для построения защищенных квантовых линий связи;

4) квантовые сенсоры и датчики (*Quantum Sensing*), обеспечивающие высокую точность при малых размерах и предназначенные, в частности, для медицинской диагностики, систем безопасности, а также отслеживания процессов старения материалов и оценки надежности конструкций;

5) квантовая метрология (*Quantum Metrology*), включающая создание сверхточных хронометров, существенно превосходящих по своим характеристи-

кам атомные часы. Это позволит значительно улучшить качество геопозиционирования (до миллиметровой погрешности) за счет улучшения синхронизации сигналов спутников.

Вместе с тем развитие КТ до настоящего времени остается в тени более известных и коммерчески прибыльных научно-технических направлений (био-, нано- и информационные технологии, энергосбережение и нетрадиционная энергетика, робототехника, фармацевтика). Большинство публикаций, в которых проводится анализ работ в сфере КТ, выполнены в научно-популярном жанре или представляют собой плохо согласующиеся экспертные мнения и прогнозы. Практически отсутствуют комплексные наукометрические исследования, позволяющие на основе авторитетной библиографической базы получить более объективные оценки развития данного научного направления в мире.

Этой актуальной проблематике посвящена представленная работа, в которой с позиции наукометрического подхода проводится анализ публикационной активности различных стран по КТ, выявляются занимаемые ими позиции по основным разделам квантовых технологий, изучаются вопросы международного сотрудничества. Таким образом, в работе проводятся исследования на макроуровне, на котором объектами изучения являются страны и ключевые направления развития КТ. Анализ на микроуровне (оценка показателей организаций, ученых, выявление профильных журналов и конференций) в большей степени доступен специалистам и при необходимости может быть реализован через возможности широко известных электронных ресурсов – *Google Scholar, Microsoft Academic Search и eLibrary*.

Результаты статьи, как представляется, позволят ученым, аспирантам, инженерам, экспертам и чиновникам более точно и выверено формулировать стратегию развития КТ в России, прогнозируя появление новых точек роста и технологических прорывов.

МЕТОДИКА ИССЛЕДОВАНИЙ

Несмотря на достаточно длительное развитие (в частности, в рамках квантовой физики), в настоящее время предметная область КТ находится в стадии становления и характеризуется постепенной трансформацией теоретического задела в прикладные разработки. Вследствие интенсивных научных изысканий появляются новые направления исследований, которые существенным образом дополняют словарь терминов рассматриваемой предметной области, вводят дополнительные понятия и определения. Автор разделяет мнение ряда экспертов, которые относят КТ к формирующимся научным направлениям, находящимся в стадии становления («*emerging technology*»). Это соответствует классификации общеевропейского флагманского проекта «*Quantum Technologies Flagship*», в котором работы по созданию квантовых устройств запланированы в рамках программы «Будущих и появляющихся технологий» (*Future and Emerging Technologie*s) [1].

Анализ слабоструктурированных областей является сложной и нетривиальной задачей. Экспертные оценки чаще всего не способны охватить весь спектр исследований. Они сложно организуемы, трудозатратны и дорогостоящи. При этом возможно лоббирование определенных научных направлений. В нашей работе для повышения достоверности результатов и выводов принято решение применить инструментарий наукометрии, позволяющий получать более объективные оценки за счет обработки и анализа профильной документальной информации, которая содержится в авторитетных международных реферативных (библиографических) базах данных (БД) таких, как *Web of Science, Scopus, Google Scholar*.

Выбор базы данных оказывает значимое влияние на получаемые результаты. В *Web of Science* проводится тщательный («осторожный») отбор информационных ресурсов, прежде всего, из числа американских журналов и конференций. Как следствие, *Web of Science* несколько завышает показатели американских авторов, а *Scopus,* ориентированный на европейские журналы, – ученых из Евросоюза. При применении *Web of Science*, как представляется, получаются более «жесткие» оценки (в частности, для российских публикаций), чем в *Scopus*. В настоящей статье принято решение проводить анализ по БД *Web of Science*.

При составлении запроса к БД *Web of Science* использовалась наиболее простая стратегия – задавалось единственное ключевое слово «*Quantum**». Далее нами слово «библиографический» опускается, поэтому, когда говорится о «статье», «публикации», «документе» имеется в виду их библиографическое описание, доступное в БД *Web of Science*. Обращение к БД *Web of Science* проводилось 10 октября 2017 г.

Для исследования взят семнадцатилетний промежуток времени с 2000 по 2016 гг. Все изучаемые публикации – англоязычные, тип материала не учитывался (статья в журнале, доклад на конференции, обзор, редакторская колонка, книга, рецензия на книгу и т.п.).

Анализировались следующие поля библиографического описания:

- название, аннотация и ключевые слова;
- место публикации (название журнала, конференции и т.п.);
- авторы и место работы авторов;
- год публикации

Для статистической обработки результатов наукометрических исследований применялась программа *STATISTICA*.

При анализе интенсивности международного сотрудничества в случае, если материал подготовлен авторами из нескольких стран, то каждой стране засчитывалось по одной публикации.

К российским статьям отнесены научные тексты, в которых хотя бы у одного автора в поле «место работы» фигурировало англоязычное название страны – «*Russia*». При расчете публикаций Великобритании учитывались все печатные труды, сделанные учеными из Англии, Северной Ирландии, Шотландии и Уэльса.

Выборка документов по КТ, полученная из БД *Web of Science* за указанный период времени, изучалась по следующим аспектам.

1. Оценка темпов ростов публикаций по КТ в ведущих странах.

2. Выявление государств-лидеров по количеству публикаций и определение областей их научного доминирования.

3. Сопоставление публикационной активности по КТ в России и государствах-лидерах.

4. Анализ интенсивности международного сотрудничества, идентификация основных участников.

РЕЗУЛЬТАТЫ АНАЛИЗА КОЛИЧЕСТВЕННЫХ ПОКАЗАТЕЛЕЙ

Оценка темпов ростов публикаций в области квантовых технологий

В период 2000-2016 гг. количество публикаций в ведущих странах по вопросам, связанным с квантовыми технологиями, возросло в два раза: 2000 г. – 15068 публикаций, 2016 г. – 31577 (рис. 1 и табл. 1), данные приводятся для первых пятнадцати стран: США, КНР, ФРГ, Япония, Великобритания, Индия, Россия, Франция, Италия, Республика Корея, Канада, Испания, Иран, Бразилия, Швейцария). Общемировое количество публикаций в мире по квантовым технологиям в 2016 г. составило 39331.

Мировые темпы роста публикаций в области КТ достаточно высоки, хотя они и уступают другим быстро развивающимся направлениям. Например, количество публикаций по тематике нетрадиционной энергетики в период с 2001 по 2013 год выросло более чем в 11 раз (запрос «*Solar Hydrogen Generation*» к БД *Web of Science*) [10]. Такие различия еще раз подтверждают целесообразность отнесения КТ к формирующимся научным направлениям (*emerging technology*). «Ошеломляющие» темпы роста КТ можно прогнозировать, если в ходе выполнения намеченных крупномасштабных программ в США, КНР и ЕС будут получены инновационные решения с высоким потенциалом практического использования, т.е. рассматриваемое научное направление перейдет в стадию сформировавшихся устойчиво развивающихся промышленно-ориентированных технологий (*advanced technology*).

На национальном уровне наибольшие темпы роста числа печатных работ за рассматриваемый семнадцатилетний период отмечаются в Китае – 7,67. Затем следует Индия – 6,17, Республика Корея – 4,15, Канада – 2,7, Бразилия – 2,46. США, Германия, Япония, Великобритания, Франция и Россия имеют весьма умеренные темпы роста равные соответственно 1,62; 1,57; 1,28; 1,83; 1,86 и 1,52.

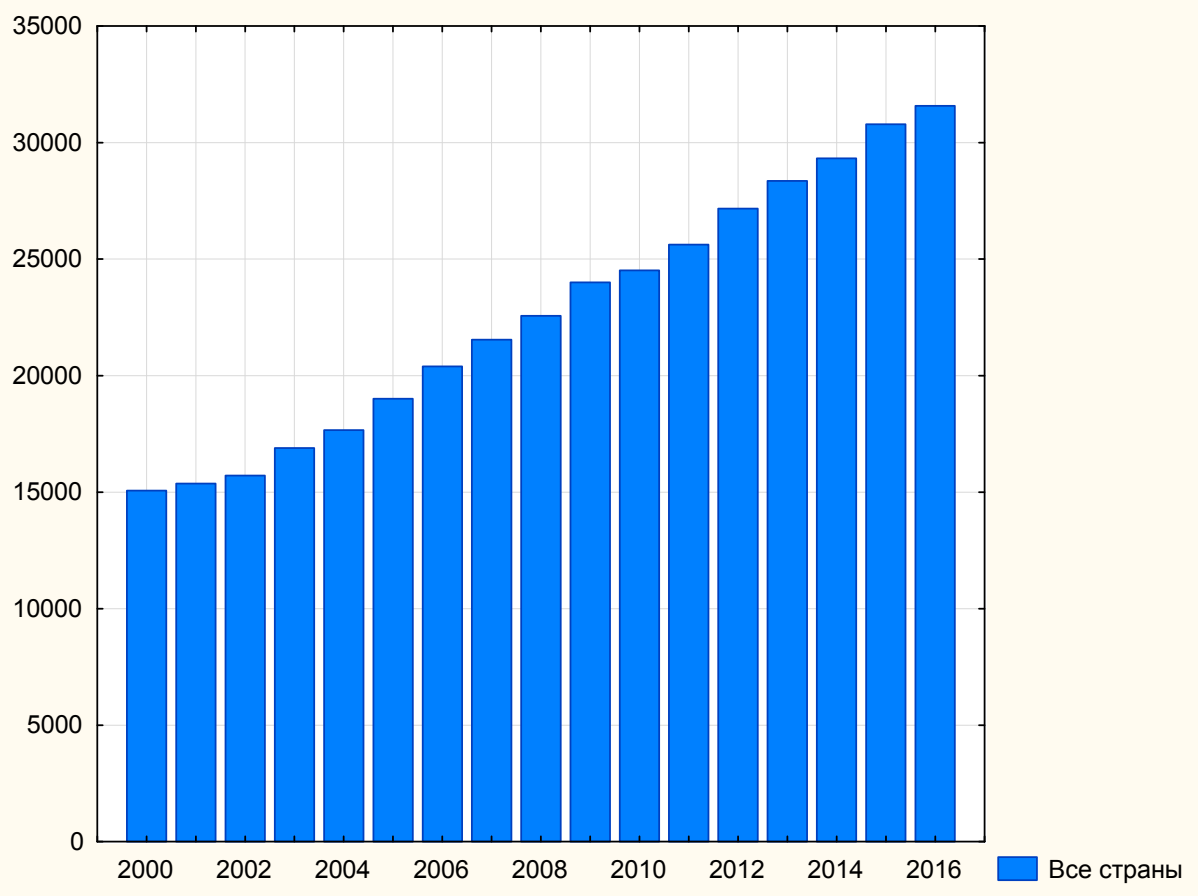


Рис. 1. Изменение числа публикаций по квантовым технологиям за 2000-2016 гг.

Таблица 1

Количество публикаций ведущих стран в области КТ за период 2000-2016 гг.

	Все страны	США	КНР	ФРГ	Япония	Великобритания	Индия	РФ	Франция
2000	15068	5069	1233	2583	1986	1385	410	1537	1215
2001	15368	4826	1371	2516	2250	1356	461	1592	1264
2002	15712	5166	1455	2664	2017	1308	494	1565	1314
2003	16899	5604	1604	2610	2189	1496	493	1570	1383
2004	17662	5779	1940	2675	2058	1509	510	1646	1494
2005	19014	6290	2377	2823	2274	1556	611	1501	1521
2006	20394	6552	3040	2975	2258	1677	722	1624	1727
2007	21542	6634	3483	3048	2457	1815	795	1621	1813
2008	22569	6759	4071	3173	2393	1960	849	1628	1861
2009	24006	7214	4476	3385	2553	1999	1070	1679	1945
2010	24521	7282	4879	3629	2403	1951	1176	1646	1998
2011	25623	7313	5611	3626	2461	1930	1379	1740	1966
2012	27174	7824	6457	3761	2460	2098	1674	1815	2025
2013	28357	7997	7156	3782	2603	2070	1839	1831	2068
2014	29324	7927	7990	3844	2343	2348	2124	1961	2092
2015	30790	7934	8972	4038	2404	2490	2394	2279	2223
2016	31577	8215	9463	4068	2559	2546	2532	2340	2265

Выявление стран-лидеров по количеству публикаций

Первое место в 2016 г. по количеству публикаций занял Китай – 9463 статьи (почти 30 % от общего числа печатных работ), второе место у США – 8215 (26 %), третье место у Германии – 4068 (12,8 %). По количеству проиндексированных документов в *Web of Science* в 2016 г. лидирующие позиции также у Японии – 2559 (8,1 %), Великобритании – 2546 (8 %), Индии – 2532 (8 %), России – 2340 (7,41 %), Франции – 2265 (7,17 %). Остальные страны в 2016 г. имели менее двух тысяч работ.

Изменение числа публикаций за семнадцатилетний период для ряда стран (США, КНР, ФРГ, Япония, Великобритания, Россия) приведены на рис. 2.

Несмотря на лидерство Китая по количеству публикаций в 2016 г., за весь семнадцатилетний период наибольшее число статей было опубликовано в США – 114385. Затем следуют Китай – 75576, Германия – 55200, Япония – 39669, Великобритания – 31494, Франция – 30174, Россия – 29575, Индия – 19533.

Данные табл. 1 и зависимости, представленные на рис. 2, показывают чрезвычайно быстрый ежегодный рост публикаций в КНР.

На момент написания статьи нами был сделан прогноз: ожидаемое в 2017 г. число американских и китайских работ. Для этого построена линейная парная регрессия: $Y = b_0+b_1*X$, где: Y – количество публикаций, X – годы, b_0 и b_1 – неизвестные коэффициенты парной линейной регрессии.

Определение оценок коэффициентов b_0 и b_1 затруднялось из-за малого размера выборки (всего семнадцать отчетов) и невыполнения исходных предположений регрессионного анализа. Оценивание по малым выборкам, как известно, может привести к получению неадекватных моделей. Принимая во внимание существование этих негативных факторов, мы, тем не менее, определили оценки неизвестных коэффициентов с помощью программы *STATISTICA*.

Линейная регрессия хорошо аппроксимирует зависимость для США: $Y_{(США)} = 4770,9+217,5X$. Коэффициент детерминации $R^2=0,98$, оба коэффициента b_0 и b_1 значимы. Прогноз на 2017 год – 8686 публикаций (доверительный интервал [8451; 8920], уровень значимости $\alpha=0,05$).

Предварительный анализ зависимости ежегодного роста публикаций в КНР позволил выбрать два варианта построения регрессионных зависимостей.

1 вариант. Парная линейная регрессия

$$Y = b_0+b_1*X.$$

2 вариант. Полиномиальная регрессия второго порядка

$$Y = b0+b1*X + b1*X^2.$$

1) Линейная регрессия имеет вид:

$$Y_{(КНР)1} = 961,19+630,68X.$$

Коэффициент детерминации $R^2=0,99$, оба коэффициента b_0 и b_1 значимы. Прогноз на 2017 год – 9795 публикаций (доверительный интервал [9405; 10185], уровень значимости $\alpha=0,05$).

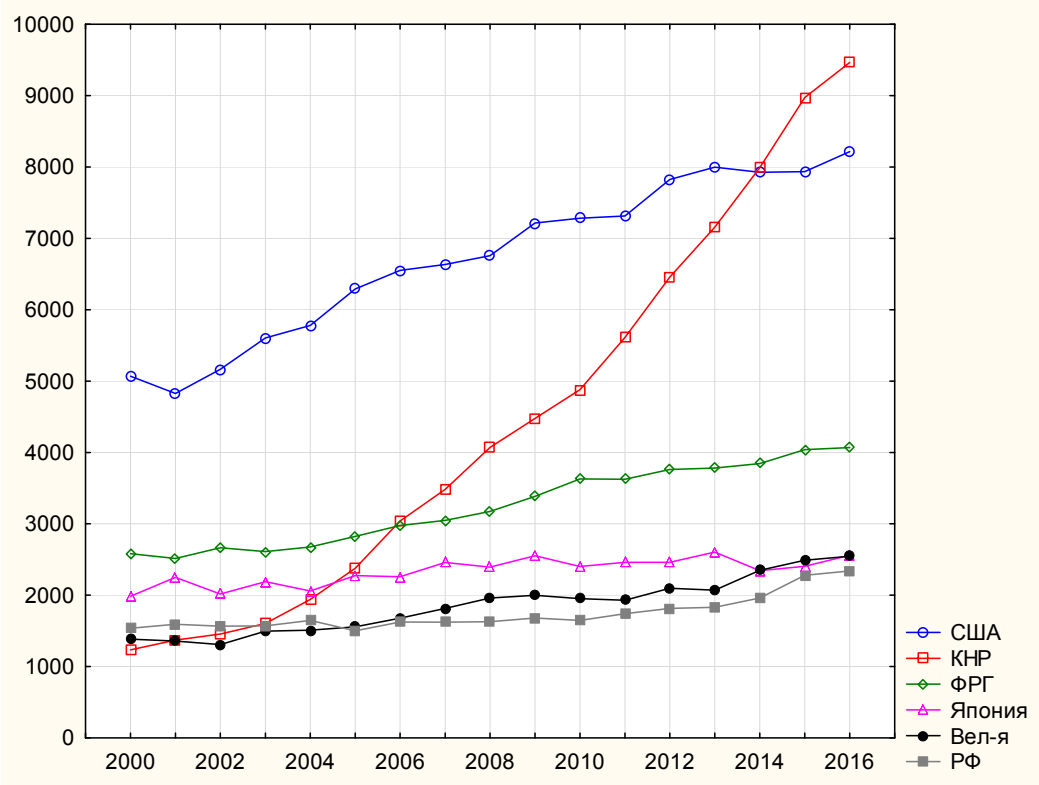


Рис. 2. Количество публикаций по ведущим странам за 2000-2016 гг.

2) Полиномиальная регрессия второго порядка имеет вид: $Y_{(КНР)2} = 967+111,95X+23,52X^2$. Коэффициент детерминации $R^2=0,99$ оба коэффициента значимы при $\alpha=0,05$. Прогноз на 2017 год – 1060 публикаций. Важно отметить, что данное значение попало в доверительный интервал для линейной парной регрессии, т.е. на незначительном промежутке времени линейная и полиномиальная регрессии дали близкие прогнозы.

Как отмечалось ранее, все вышеописанные исследования проводились на основе сведений, полученных при обращении к БД *Web of Science* 10 октября 2017 г. Достаточно длительный период подготовки материала позволил автору провести проверку сделанных прогнозов. Для этого 5 апреля 2018 г. было выполнено еще одно обращение к БД *Web of Science*, в результате которого получены данные за 2017 г.

В США и КНР в 2017 г. происходили разнонаправленные процессы. В Китае отмечался беспрецедентный рост научных трудов по квантовым технологиям, китайские ученые за прошлый год опубликовали 10546 работ (на 1083 больше чем в 2016 г.). В США общее число статей составило 8140 (уменьшение на 75 по сравнению с 2016 г.). Полученные показатели публикационной активности (очень быстрый рост в Китае и небольшой спад в США) существенно выделяются среди наблюдений за несколько последних лет.

К сожалению, построенные регрессионные модели не смогли адаптироваться к новым данным и предсказать подобные результаты. Как следствие, получены недостоверные прогнозы (полиномиальная регрессия второго порядка существенно «недооценила» число публикаций в Китае, результат линейной регрессии для США не попал в доверительный интервал).

На взгляд автора, «провал» математического моделирования объясняется не столько «наивностью» подхода к получению оценок на «малых» выборках, сколько является результатом действия факторов, которые практически не измеряемы, их сложно включить и учесть в модели (некоторые из них способствовали увеличению нелинейного роста публикаций в Китае и замедлили аналогичный процесс в США). Возможно, одним из таких неучтенных (но крайне важных) факторов является усиление финансирования китайской науки и, как следствие, стремительное увеличение числа ученых и количества издаваемых работ по перспективным направлениям исследований. В тоже время в США Администрация Трампа сокращает бюджетные ассигнования на гражданские НИОКР, концентрируя усилия на военных разработках. Нельзя исключить (и это мнение разделяет часть опрошенных отечественных экспертов по КТ), что некоторые результаты в области квантовых технологий будут издаваться американцами в ограниченном количестве из-за их особой значимости при создании современного вооружения, военной и специальной техники.

На основе представленных выше наукометрических данных можно сделать вывод о том, что Китай в ближайшие годы упрочит свой отрыв от других стран по количеству документов по КТ, ежегодно индексируемых в БД *Web of Science,* и при сохранении имеющихся темпов роста к 2030 г. догонит США по общему числу публикаций.

Распределение статей по разделам рубрикатора Web of Science

Анализ публикаций по КТ показывает, что основное число научных трудов соответствует следующим двенадцати разделам рубрикатора *Web of Science* (табл. 2):

1) прикладная физика (ПФ – *Physics Applied*);

2) междисциплинарные вопросы материаловедения (МВМ – *Materials Science Multidisciplinary*);

3) физическая химия (ФХ – *Chemistry Physical*);

4) оптика (ОП – *Optics*);

5) междисциплинарные вопросы физики (МВФ – *Physics Multidisciplinary*);

6) физика конденсированных сред (ФКС – *Physics Condensed Matter*);

7) общие вопросы химии (ОВХ – *Chemistry Multidisciplinary*);

8) атомная, молекулярная и химическая физика (АМиХФ – *Physics Atomic Molecular and Chemical*);

9) нанонаука и нанотехнологии (НиН – *Nanoscience and Nanotechnology*);

10) электротехника и электроника (ЭиЭ – *Engineering Electrical and Electronic*);

11) физика квантовых полей (ФКП – *Physics, Particles & Fields*);

12) математическая физика (МФ – *Physics, Mathematical*).

В табл. 2 приведены семь стран, которые входят в первую десятку по всем указанным областям. Лидерство в каждой области по количеству публикаций принадлежит США или КНР. США занимает первое место в семи тематиках, Китай – в пяти. В десятку лучших по публикациям также достаточно часто входили Италия – 11 раз, Россия – 10 (отсутствует в междисциплинарных вопросах материаловедения и атомной, молекулярной и химической физики), а также Республика Корея, которая специализируется в пяти направлениях.

В этой табл. 2 номер столбца соответствует номеру рубрикатора *Web of Science* из вышеприведенного списка. Отметим, что каждая статья в БД *Web of Science* может быть отнесена одновременно к нескольким разделам рубрикатора, в данном исследовании учитывался только раздел рубрикатора, который был указан первым.

Проанализируем более детально показатель публикационной активности России. Насколько этот показатель согласуется с работами, издаваемыми российскими учеными в других предметных областях. Согласно результатам наукометрических исследований в области нанотехнологий, получавших в течение двух последних десятилетий самое высокое государственно-частное финансирование, за период 1990-2012 гг. россиянами было подготовлено 33538 научных документов, проиндексированных в БД *Web of Science* [11]. Несмотря на некоторое несовпадение временных периодов, сравним общее число публикаций в нанотехнологиях – 33538 и КТ – 29575. Из результатов сравнения следует, что публикационная активность российских ученых в области КТ весьма высока и близка к показателю нанотехнологической области, которая долгое время пользовалась наивысшей поддержкой государства.

Сопоставим результаты лидеров (США, Китая) и России по десяти из двенадцати вышеуказанных тематик (выбраны тематики, по которым Российская Федерация входит в первую десятку). Эти результаты представлены в табл. 3 и на рис. 3.

Данные, приведенные в табл. 3 и на рис. 3, наглядно иллюстрируют развернувшуюся борьбу за лидерство между США и Китаем. Они также показывают достаточно высокие публикационные показатели российских ученых в области КТ. Это, как представляется, свидетельствует о существенном заделе по квантовой проблематике, накопленном еще с советских времен. Большинство таких достижений сосредоточено в традиционных сферах, в которых российские ученые наиболее успешны (физика, химия).

Таблица 2

Страны, имеющие наибольшее количество публикаций по разделам рубрикатора *Web of Science*

	1	2	3	4	5	6	7	8	9	10	11	12
США	1342	1215	1380	1208	910	1090	793	1098	1291	563	613	356
КНР	1624	1949	1231	1143	1212	843	1518	704	891	428	254	321
ФРГ	640	427	508	624	499	687	377	573	287	238	297	233
Великобритания	350	244	257	349	337	256	189	345	162	174	227	164
Франция	328	253	303	316	279	355	199	287	188	146	157	151
Япония	562	381	312	296	306	301	246	257	215	257	177	128
Индия	353	424	417	216	207	306	338	232	229	247	144	97

Таблица 3

Количество публикаций США, КНР и России по десяти разделам рубрикатора *WOS* в 2016 г.*

	США	КНР	РФ
ПФ	1342	1624	342
ФХ	1380	1231	237
ОП	1208	1143	352
МВФ	910	1212	334
ФКС	1090	843	354
ОВХ	793	1518	179
НиН	891	1291	142
ЭиЭ	563	428	129
ФКП	613	254	202
МФ	356	321	148

* Для обозначения предметных областей использованы введенные ранее сокращения.

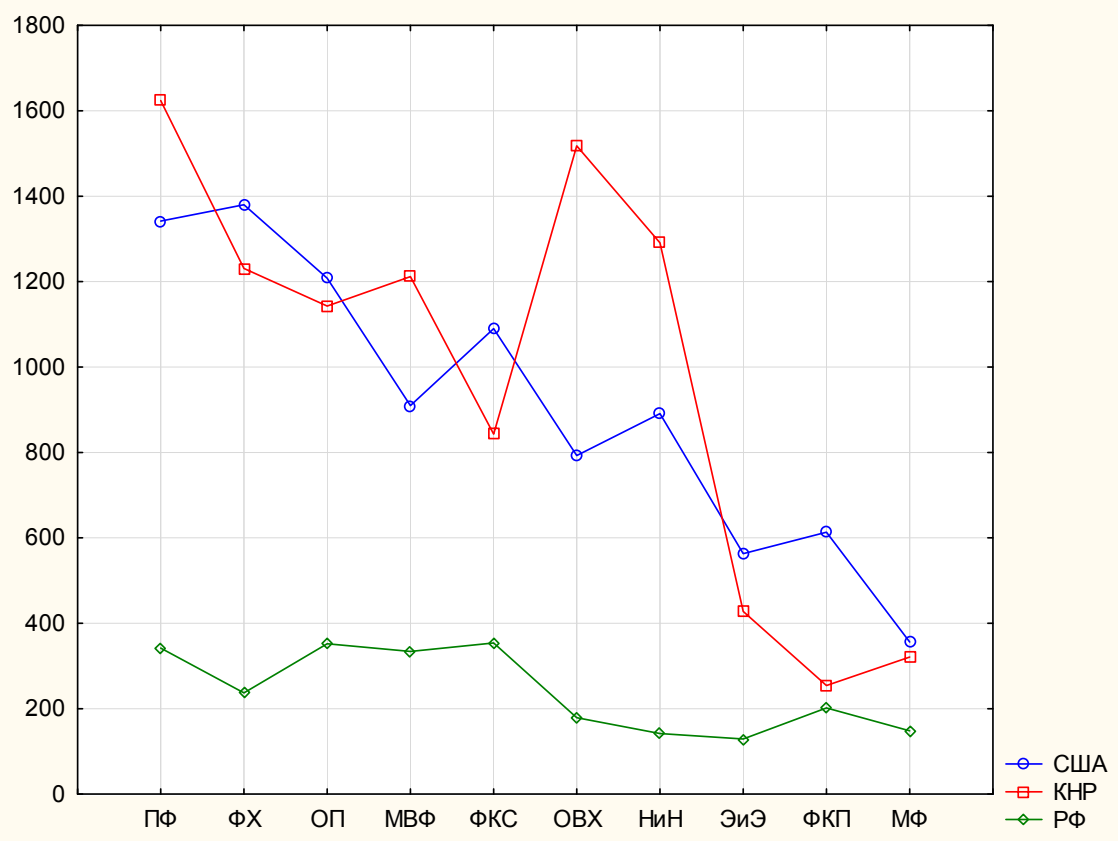


Рис. 3. Количество публикаций США, КНР и России по десяти разделам рубрикатора WOS в 2016 г.

Анализ числа публикаций в мире и ведущих странах по тематике «квантовые вычисления» (Quantum Computing)

Для квантовых технологий, также как и для большинства других научных направлений, отсутствует единый общепризнанный рубрикатор. Рубрикаторы БД *Web of Science, Scopus, Google Scholar* существенно различаются. В ряде случаев эти тематические разделы медленно обновляются и не в полной мере описывают быстро развивающиеся предметные области. В связи с этим представляет интерес наукометрический анализ тематик, которые наиболее широко обсуждаются в научном сообществе и в них ожидаются значительные технологические прорывы. К одной из таких тематик относятся работы по соз-

данию универсального квантового компьютера и квантовых симуляторов. Для выявления тенденций развития данного направления зададим в *Web of Science* в качестве ключевых слов «*Quantum Computing*» (квантовые вычисления). Результирующая выборка будет содержать это словосочетание (в строго заданной последовательности) во всех полях библиографического описания.

Полученные результаты представлены в табл. 4, а на рис. 4 и 5 проиллюстрировано изменение общего числа публикаций за анализируемый период и приведены показатели четырех стран-лидеров: США, Китая, Канады и Японии.

Сравним результаты табл. 2 и 4. Необходимо отметить наличие ряда существенных различий в случае анализа более «узкой» тематики, которая определяется не по разделам рубрикатора БД *Web of Science*, а соответствует приоритетным направлениям исследований крупных национальных программ.

Отметим следующие особенности результатов, которые приведены в табл. 4. Во-первых, число публикаций по тематике квантовых вычислений за рассматриваемый период увеличилось практически в 4 раза, т.е. росло значительно быстрее количества статей по квантовым технологиям в целом. Во-вторых, лидеры остались неизменными. США и Китай существенно опережают другие страны. Однако в данной проблематике наблюдается достаточно очевидное доминирование США, которые прочно удерживают передовые позиции в разработке квантовых симуляторов и компьютеров. В-третьих, в борьбе за третье место появилась Канада, отсутствовавшая в девяти из двенадцати позиций рубрикатора *Web of Science* в первой десятке (хотя ранее нами был отмечен достаточно быстрый рост в период 2000-2016 гг. числа канадских статей по КТ – 2,7). Это можно объяснить важной ролью, которую в настоящее время играет канадская фирма *D-Wave Systems* в развитии квантовых вычислений. Все первое десятилетие XXI в. *D-Wave Systems* получала значительные венчурные инвестиции и практически не имела соперников в области построения квантовых симуляторов.

Особый интерес для анализа представляет резкий всплеск активности, который наблюдается в 2009 г. на всех зависимостях (см. рис. 4 и 5). Возможно, это объясняется высоким интересом научного сообщества к проблематике создания квантовых симуляторов после проведения конференции по суперкомпьютерам (*Supercomputing-SC07*) в США в ноябре 2007 г.. На этой конференции *D-Wave Systems* провела онлайновую демонстрацию 28-кубитного квантового компьютера для идентификации изображений и заявила о выпуске в 2008 г. вычислительного устройства с 512 кубитами [12]. Проведенная демонстрация спровоцировала оживленную дискуссию среди специалистов, причем преобладало критическое и скептическое мнение о работоспособности созданного варианта.

Таблица 4

Количество публикаций по квантовым вычислениям в мире и ведущих странах

	Общее число статей	**США**	**КНР**	**Канада**	**Япония**
2000	118	54	4	4	5
2001	168	71	6	6	10
2002	203	80	14	14	19
2003	246	101	14	18	30
2004	214	85	10	21	17
2005	269	107	22	18	15
2006	233	87	22	16	14
2007	312	98	41	23	20
2008	313	97	41	14	18
2009	516	161	89	58	37
2010	271	95	48	20	23
2011	283	98	40	22	22
2012	298	87	44	21	27
2013	298	96	47	22	21
2014	340	96	70	22	26
2015	380	110	65	17	20
2016	407	135	70	27	25

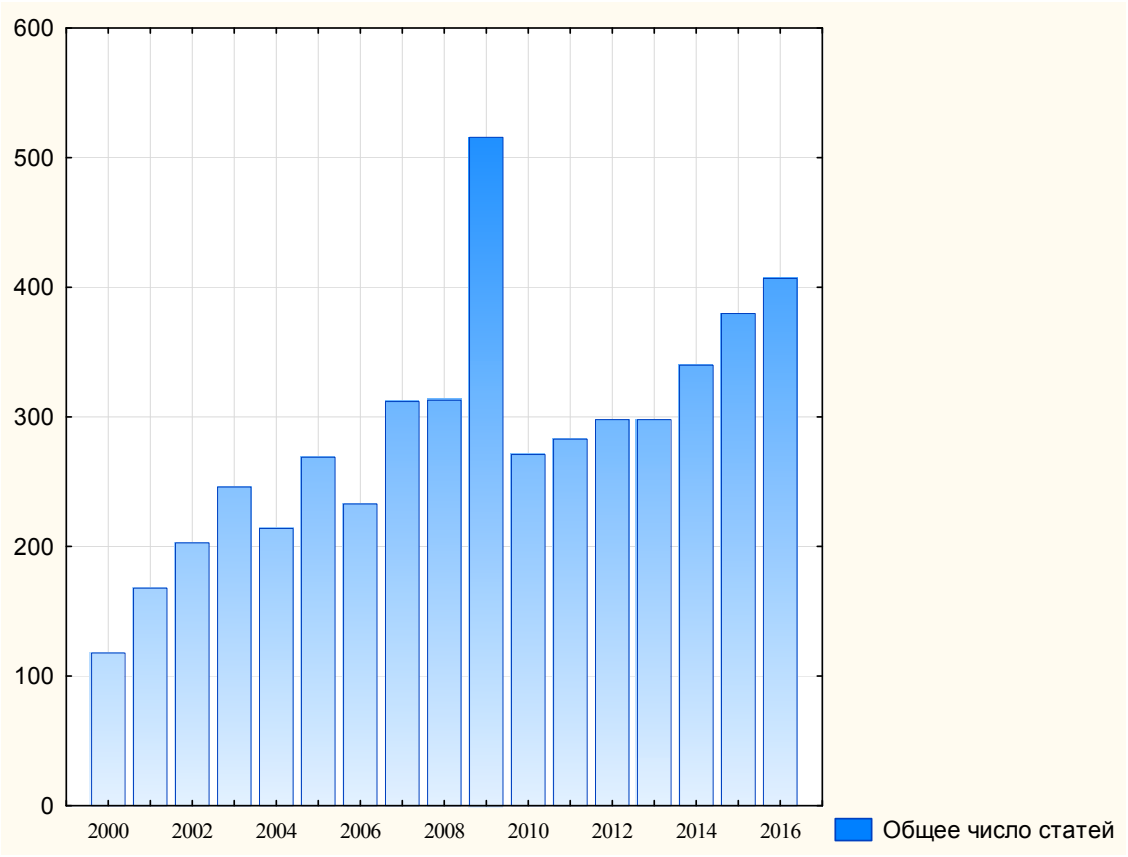


Рис. 4. Изменение числа публикаций по квантовым вычислениям в период 2000-2016 гг.

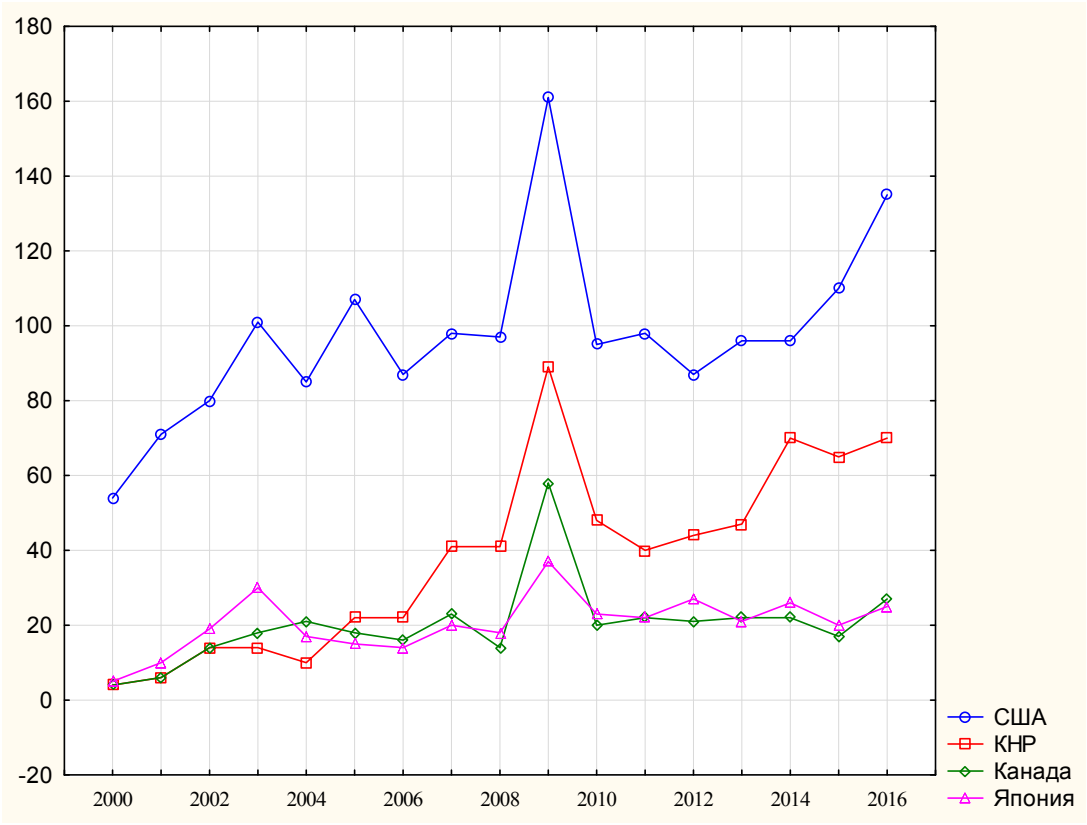


Рис. 5. Количество публикаций ведущих стран по квантовым вычислениям в период 2000-2016 гг.

Негативная реакция экспертов во многом объясняет спад интереса и выход на «плато» зависимостей, приведенных на рис. 5, когда число публикаций в период 2010-2014 гг. практически не изменялось. Заметный рост начался лишь в 2014 г. в США и во многом обусловлен существенными результатами, полученными как в американских университетах и стартапах, так и в крупных корпорациях, прежде всего *Google, Intel, IBM, Microsoft* [13].

В связи с достаточно очевидным доминированием США в области квантовых вычислений возникает актуальный вопрос – также ли существенно лидерство США по другим тематикам (например, по двенадцати разделам рубрикатора БД *Web of Science*, которые рассматривались ранее). Напомним, что согласно табл. 2 США занимает первое место в семи тематиках, Китай – в пяти. Для сопоставления публикационной активности США и КНР (обнаружения значимых различий) применим непараметрический критерий знаков и критерий Вилкоксона для связанных пар наблюдений. Проверим нулевую гипотезу об однородности генеральных совокупностей по попарно связанным выборкам. Данными для анализа являются первые две строки табл. 2, в качестве вычислительного средства используется программа *STATISTICA*, уровень значимости α=0,05.

Оба теста не обнаружили существенных различий между публикационными показателями США и КНР по двенадцати тематикам рубрикатора БД *Web of Science*. Нет оснований по имеющимся данным отвергать нулевую гипотезу (выборки однородны, их элементы взаимозаменяемы).

Вместе с тем в наиболее перспективной области КТ – квантовых вычислениях, развитие которой способно привести к революционным изменениям, США сохраняет безоговорочное публикационное лидерство. Это дает возможность предположить, что в случае, если универсальный квантовый компьютер все-таки появится, то первыми его создадут именно американские специалисты.

Анализ интенсивности международного сотрудничества, выявление основных участников

В настоящее время в условиях глобализации научных исследований существенно возрастает число совместных публикаций ученых из различных стран. Особенно такое взаимодействие сильно в областях, которые находятся в стадии становления и требуют значительных инвестиций для реализации научных прорывов. Поэтому софинансирование НИР несколькими государствами является одним из широко распространенных способов осуществления крупных проектов и обуславливает появление большого числа межнациональных публикаций. Еще одним важным фактором является возрастающая мобильность ученых, их стремление работать в передовых научных центрах и университетах.

В табл. 5 представлено количество межнациональных работ за рассматриваемый период времени. На главной диагонали стоит общее количество национальных публикаций, в остальных ячейках содержится число научных материалов, в подготовке которых принимали участие, как минимум, ученые из двух государств, название которых указаны в заголовке строки и столбца (в таблице размера 11x11 приведены результаты только для одиннадцати ведущих стран). Как справедливо отмечается специалистами в области наукометрии, публикации, подготовленные в ходе международного сотрудничества, чаще всего имеют более высокий показатель цитирования и издаются в наиболее рейтинговых журналах [14, 15].

Несмотря на сильную конкуренцию между США и КНР практически по всем направлениям развития КТ, борьба за лидерство сопровождается высоким уровнем взаимодействия. Так, США является основным партнером КНР по изданию совместных публикаций (затем следуют с большим отрывом Германия и Япония). Что касается США, то для них основным партнером является Германия. Китай находится на втором месте и по количеству совместных работ лишь незначительно уступает ФРГ.

Для остальных стран, представленных в табл. 5 и на рис. 6, США также является ключевым партнером. Именно американцы имеют наибольшее число статей, выполненных в соавторстве с иностранцами. Это объясняется несколькими факторами: лидерством американских ученых, объединяющих вокруг своих исследований специалистов из других стран, многие из которых временно работают в США или находятся на стажировке; наличием в США большого числа высокоцитируемых авторитетных журналов, составляющих основу *Web of Science*; высоким финансированием КТ; наиболее совершенной лабораторной базой и информационно-телекоммуникационной инфраструктурой. Все это в совокупности и образует интеллектуальную среду, формирующую новые теоретические результаты и инновационные решения.

Анализируя российское международное сотрудничество, отметим, что большинство совместных публикаций выполняется в кооперации с немецкими и американскими учеными. Ранее проведенные наукометрические исследования в других научных областях показывают, что по всему фронту исследований именно страны Евросоюза, прежде всего, Германия, Франция и Великобритания, являются основными партнерами России при издании совместных статей, причем с европейскими учеными выпускается больше работ, чем с американцами [11, 14, 15].

Особую озабоченность вызывает мизерная кооперация со странами БРИКС (совокупное число работ с Китаем, Индией и Бразилией в четыре раза меньше, чем российско-германских публикаций). Россия крайне слабо участвует также в международном сотрудничестве на азиатском направлении (общее число статей с Японией, Китаем, Южной Кореей и Индией составляет 1647 публикации – меньше общего числа российско-французских работ). На аналогичную ситуацию обратили внимание авторы, проводившие наукометрические исследования в области нанотехнологий [11].

Таблица 5

Количество работ по квантовым технологиям, выполненным в рамках международного сотрудничества

	США	КНР	ФРГ	Япония	Великобритания	Франция	Россия	Индия	Республика Корея	Бразилия	Нидерланды
США	114385	7056	7804	3870	4608	3951	2899	1328	2152	1150	1376
КНР	7056	75576	1937	1517	1106	733	389	274	753	221	264
ФРГ	7804	1937	55200	1770	3685	3602	4178	805	548	722	1275
Япония	3870	1517	1770	39669	1330	1186	807	395	754	147	357
Велико-британия	4608	1106	3685	1330	31494	2246	1186	466	427	501	774
Франция	3951	733	3602	1186	2246	30174	1796	448	304	647	694
Россия	2899	389	4178	807	1186	1796	29575	178	273	443	465
Индия	1328	274	805	395	466	448	178	19533	474	192	103
Ю.Корея	2152	753	548	754	427	304	273	474	14485	103	68
Бразилия	1150	221	722	147	501	647	443	192	103	10896	115
Нидерланды	1376	264	1275	357	774	694	465	103	68	115	7736

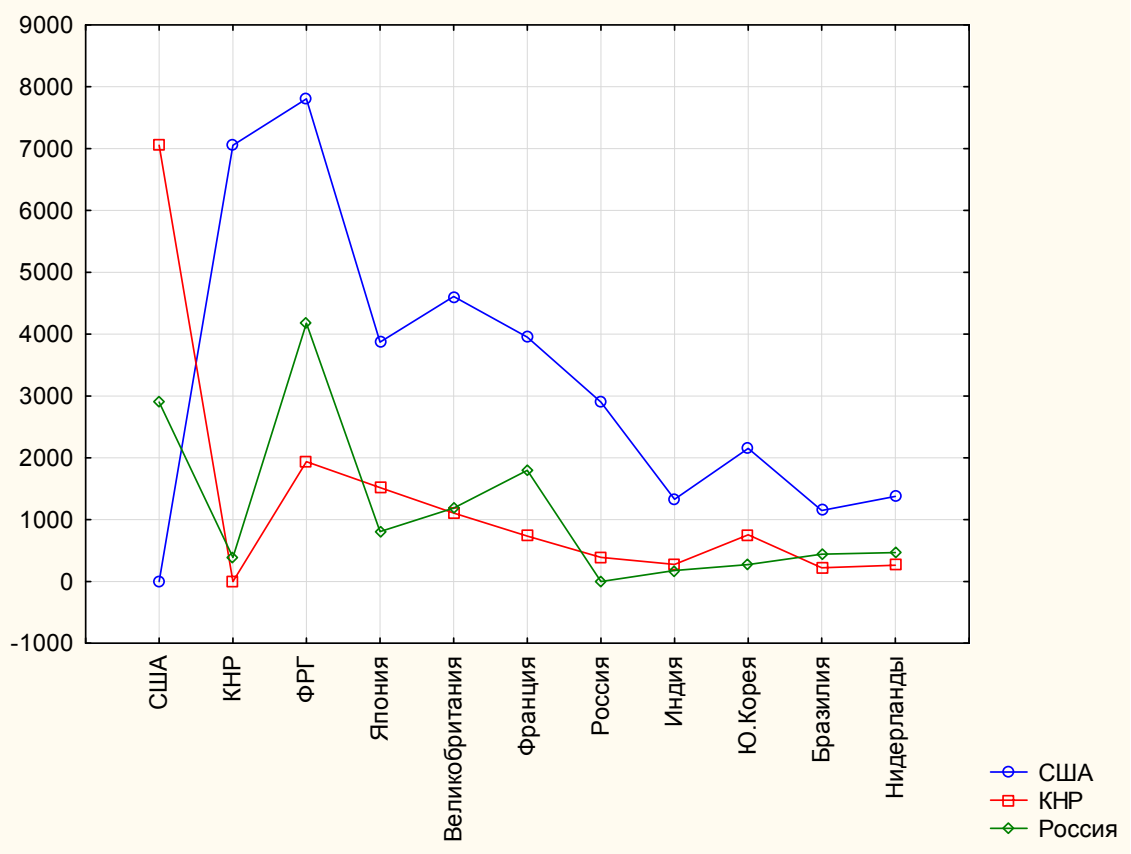


Рис. 6. Количество публикаций ведущих стран, выполненных в рамках международного сотрудничества

Вместе с тем, как отмечалось ранее, именно в азиатских странах, прежде всего Китае, Индии и Республике Корея, наблюдаются самые высокие темпы роста публикационной активности, которая отражает формирование в этих государствах значительного научного задела в области КТ. Недооценка данного направления сотрудничества может весьма негативно сказаться на качестве российских НИР по квантовым тематикам, особенно в современных условиях прекращения в связи с санкциями ряда международных проектов с участием России и западных стран.

Количество российских публикаций в области КТ, подготовленных в сотрудничестве с иностранцами, составляет 12614 статей (42,6 процентов от общего массива российских работ, проиндексированных в *Web of Science* в период с 2000 по 2016 гг.). При этом аналогичный показатель для россиян, например, в сфере биотехнологий (бионаук, *Biosciences*) составляет 64,7% в период с 2009-2013 гг. [14]. Таким образом, наблюдается определенное отставание по числу межнациональных публикаций, подготовленных с участием россиян. Причем важно отметить, что показатель по КТ является завышенным, так как в случае наличия у статьи четырех авторов, например из России, Германии, Франции и Великобритании, эта работа будет учтена в нашем исследовании трижды при оценке российско-немецкого, российско-французского и российско-английского сотрудничества.

ЗАКЛЮЧЕНИЕ

В настоящей работе проведен комплексный наукометрический анализ одного из самых перспективных быстро развивающихся научных направлений – квантовые технологии. Результаты исследований позволяют сделать следующие выводы.

1. Изучаемая предметная область имеет высокие темпы роста и находится в стадии быстрого развития. За рассматриваемый период времени общее число профильных публикаций увеличилось более чем в два раза. Прогнозируется продолжение данной тенденции благодаря активному государственно-частному финансированию НИОКР.

2. Как и во многих других перспективных технологиях, в квантовых наблюдается соперничество между двумя лидерами – США и КНР. При этом США сохраняет первенство по общему количеству публикаций, а КНР – по темпам роста публикационной активности и количеству ежегодно издаваемых научных материалов.

3. По ряду важных направлений работ в сфере квантовых технологий Россия имеет достаточно высокие количественные показатели и входит в первую десятку стран. В международном сотрудничестве по квантовым технологиям основными партнерами российских ученых являются специалисты из Германии, США, Франции и Великобритании (при крайне низком взаимодействии с крупнейшими азиатскими странами – Китаем, Японией, Индией и Республикой Корея). Это, в целом, соответствует результатам аналогичных наукометрических исследований, в которых изучалось международное научно-техническое сотрудничество России в других предметных областях (например, нано- и биотехнологиях [11, 14]).

Несмотря на высокую объективность и достоверность наукометрических оценок, представляется целесообразным указать возможные причины, которые обуславливают некоторое искажение результатов.

Первая причина – неполнота исходных данных. В частности, у нас не было сведений о ежегодном государственном и частном финансировании исследований по КТ в разных странах, количестве научных сотрудников, задействованных в исследованиях и их возрастной структуре, образовательных программах по подготовке специалистов в сфере КТ. Все это затрудняет прогнозирование темпов роста публикационной активности национальных научных организаций. При работе над статьей не удалось также получить доступ к данным по числу ежемесячных обращений к поисковым системам *Google* и Яндекс с запросом, содержащим термин «*Quantum*».

Рассматриваемые статьи были ограничены только англоязычными работами. Российские публикации, не вошедшие в международную БД *Web of Science*, но содержащиеся в русскоязычной цифровой научной библиотеке *eLibrary* отдельно не рассматривались.

Вторая причина – недостатки наукометрического подхода, которые широко известны и активно обсуждаются в научном сообществе (см., например [16]), но в данной работе дополнительно рассматриваться не будут.

Для нивелирования возможных недочетов, на взгляд автора, требуется обсуждение представленных результатов специалистами-предметниками, способными дополнить полученные наукометрические оценки экспертными мнениями и комментариями. Несомненно, перспективным направлением дальнейшего изучения данной предметной области является обнаружение новых трендов, зарождающихся в области КТ, на основе анализа изменения терминологии предметной области, в частности выделение так называемых слабых сигналов (*Weak Signals*) [17]. Такие исследования требуют применения более сложных методов интеллектуального анализа данных (*Data and Text Mining*), в ряде работ этот подход называется «*Tech Mining*» – выявление зарождающихся технологий (по аналогии с *Data and Text Mining*) [18].

Выводы, сделанные по результатам проведенного наукометрического анализа, представляют интерес для нескольких категорий специалистов: лиц, принимающих решения в области научной политики и ответственных за рациональную и эффективную организацию НИОКР в сфере наиболее перспективных тематик; ученых, специализирующихся в квантовых технологиях, отслеживающих тенденции их развития и нацеленных на выполнение прорывных работ по самым актуальным направлениям исследований (точкам роста в области КТ).

По мнению автора, проанализированная в работе предметная область, находится на переднем фронте научных исследований и совместно с прорывами в области глубокого машинного обучения (*Deep Learning*) и высокоточного моделировании работы человеческого мозга (*Brain-Computer Interfaces*) станет технологической основой для решения широкого спектра задач вычислительной техники, криптографии, робототехники и искусственного интеллекта,

моделирования и синтеза новых материалов, метрологии и геопозиционирования [19–22]. Квантовые технологии способны сыграть важную роль в реализации активно формирующихся концепций «Цифровой экономики», «Интернет-вещей», «Индустрии 4.0» и т.д. Главное и самое спорное в имеющихся прогнозах – сроки. По этой позиции в научном сообществе имеются существенные разногласия и приход «квантума» (появление первого универсального компьютера) ожидается в достаточно отдаленной перспективе – 2030-х гг.

СПИСОК ЛИТЕРАТУРЫ

1. Quantum manifesto. – European Commission. – 2016. – URL: http://qurope.eu/manifesto (дата обращения 05.01.2018).
2. Advancing Quantum Information Science: National Challenges and Opportunities. National Science and Technology Council. – USA. – 2016. – 16 p. – URL: https://www.whitehouse.gov/sites/whitehouse.gov/files/images/Quantum_Info_Sci_Report_2016_07_22%20final.pdf (дата обращения 05.01.2018).
3. Petta J.R. et al. Coherent Manipulation of Coupled Electron Spins in Semiconductor Quantum Dots // Science. – 2005. – Vol. 309. – P. 2180–2184.
4. Булатов И.В. В Китае открылась самая протяженная в мире квантовая коммуникационная линия. – РИА Новости. – 2016. – URL: https://ria.ru/science/20161120/1481735554.html (дата обращения 05.01.2018).
5. Квантовый компьютер D-Wave 2000Q за 15 млн долларов. – IT News. – 2017. – URL: http://information-technology.ru/news/6263-kvantovyj-kompyuter-d-wave-2000q-za-15-mln-dollarov (дата обращения 05.01.2018).
6. Chen L., Jordan S., Liu Y-K., Moody D., Peralta R., Perlner R., Smith-Ton D. Report on Post-Quantum Cryptography. – NIST. – 2016. – 15 p.
7. Минобрнауки Российской Федерации, Фонд перспективных исследований и Росатом работают над проектом по созданию квантовых вычислительных систем. – Фонд перспективных исследований – 2016. – URL: http://fpi.gov.ru/press/media/20160430 (дата обращения 05.01.2018).
8. Баулин А. Квантовая гонка: победитель получает все // Life.ru. – 2016. – URL: https://life.ru/t/технологии/407777/kvantovaia_ghonka_pobieditiel_poluchaiet_vsio (дата обращения 05.01.2018).
9. Future Directions for NSF Advanced Computing Infrastructure to Support. U.S. Science and Engineering in 2017-2020. – Washington: The National Academic press. – 2014. – 34 p.
10. Maghami M., Rezadad S., Gomes N. Qualitative and Quantitative Analysis of Solar Hydrogen Generation Literature from 2001 to 2014 // Scientometrics. – 2015. – Vol. 105. – P. 759–771.
11. Karaulova M., Gok A., Shackleton O., Shapira P. Science System Path-Dependencies and their Influences: Nanotechnology Research in Russia // Scientometrics. – 2016. – Vol.107. – P. 645–670.
12. Специалист Google представит квантовый компьютер D-Wave. – SecurityLab. – 2007. – URL: https://www.securitylab.ru/ news/307456.php (дата обращения 05.01.2018).
13. Технологические прорывы 2017 года. Когда появятся квантовые компьютеры. – Econet.ru. – 2017. – URL: https://econet.ru/articles/165528-tehnologicheskie-proryvy-2017-goda-kogda-poyavyatsya-kvantovye-kompyutery (дата обращения 05.01.2018).
14. Gureyev V.N., Mazov N., Karpenko L. Russian Bioscience Publications and Journals in International Bibliometric Databases // Serials Review. – 2015 – Vol. 41. – P. 77–84.
15. Писляков В.В. Шедевры научного творчества: анализ высокоцитируемых статей российских ученых // Научно-техническая информация. Сер. 2. – 2011. – №12. – С. 1–8.
16. Наукометрия и экспертиза в управлении наукой // Управление большими системами. Сб. трудов. Вып. 44 / под ред. Д.А. Новикова, А.И. Орлова, П.Ю. Чеботарева. – М.: ИПУ РАН, 2013. – 568 с.
17. Толчеев В.О. Автоматизированное оценивание формулировок научной новизны публикаций // Заводская лаборатория и диагностика материалов. – 2017. – Т. 83, № 5. – С. 72–78.
18. Porter A., Cunningham S. Tech Mining: Exploiting New Technologies for Competitive Advantage. – Willey, 2004 – 408 p.
19. Воронцов К. Машинное обучение: шаг в цифровую экономику. – 2017. – URL: https://www.youtube.com/watch?v=H5waFQ1ARF8 (дата обращения 05.01.2018).
20. Hu K., Chen C., Meng Q., Williams Z., Xu W. Scientific Profile of Brain-Computer Interfaces: Bibliometric Analysis in a 10-year period // Neuroscience Letters 635. – 2016. – P.61–66.
21. Львовский А. Квантовый компьютер и квантовые технологии. – Квантовый центр «Сколково». – 2016. – URL: https://www.youtube.com/watch?v=VfEstE2TuhU (дата обращения 05.01.2018).
22. Квантовые технологии – как физика сверхмалых частиц придет в нашу жизнь. – INNONECH. – 2017 – URL: http://innotechnews.com/innovations/1561-kvantovye-tekhnologii-kak-fizika-sverkhmalykh-chastits-pridet-v-nashu-zhizn (дата обращения 05.01.2018).

Материал поступил в редакцию 15.01.18.

Сведения об авторе

ТОЛЧЕЕВ Владимир Олегович – доктор технических наук, доцент, профессор кафедры управления и информатики НИУ «Московский энергетический институт»
e-mail: tolcheevvo@mail.ru

УДК 001.4: [37.018.43:004.738.5]

В.М. Московкин, Т.В. Сапрыкина

Об эволюции терминов, обозначающих дистанционное обучение, с помощью сервиса *Google Books Ngram Viewer*

Цель исследования – изучение эволюции терминов в области дистанционного, открытого и онлайнового образования и обучения. Анализ проведен с применением сервиса Google Books Ngram Viewer для выявления частоты встречаемости терминов во временном периоде по шести языковым корпусам публикаций. Определены тенденции применения терминов в сфере образования (сравниваются языковые группы), установлена взаимосвязь использования терминов в области образования с экономическими, социальными, политическими, технологическими факторами.

Ключевые слова: *образование, дистанционное образование, дистанционное обучение, дистанционные курсы, интернет-образование, онлайн-курсы, онлайн-образование, Google Books Ngram Viewer*

ВВЕДЕНИЕ

В современном мире наблюдается изменение среды и технологий процесса обучения, которые приводят к формированию новых понятий и тенденций в образовательной деятельности. Традиционное образование предусматривает постоянное взаимодействие преподавателя и студента в учебном заведении, наличие лекционных и семинарских занятий с проведением контрольных мероприятий. Однако развитие общества, инфраструктуры, а также разработка и применение новых средств коммуникаций обусловливают и изменение подходов к обучению. Изучением вопросов развития современных образовательных технологий занимаются ученые во многих странах. Одним из «новых» методов является дистанционное обучение с применением интернет-технологий.

Чем интенсивнее протекают жизненные процессы, тем всё больше интересует людей дистанционный способ получения образования. В книге «*Foundations of distance education*» *D. Keegan* пишет: «Поколение с 1970 по 2000 год является свидетелем развития всей области дистанционного обучения, которое шло параллельно с успехами и достижениями Открытого университета (*Open University*). Наблюдались значительные изменения в качестве, количестве, статусе и влиянии предоставления дистанционного образования. Это было связано с общим переходом от частных практик к государственному обеспечению данного процесса» [1, с. 3].

Следующим этапом развития дистанционного образования стало развитие сети Интернет. В своей книге «*Technology, e-learning and distance education*» *A.W.T. Bates* описывает внедрение и развитие интернет-технологий в систему обучения, в дистанционные образовательные процессы, а также показывает необходимость реструктуризации образовательных учреждений: «Дистанционное образование привело к большим изменениям в организации процесса обеспечения образования. Очевидно, разница в том, что студенты больше не обязаны посещать кампус на регулярной основе. В результате, дистанционное образование потребовало совершенно иных организационных структур, отличных от тех, которые используются в обычных учебных заведениях. Кроме того, по мере технологических изменений также возникает необходимость реорганизации учреждений для использования преимуществ новых технологий» [2, с. 17].

Некоторые исследователи выделяют факторы, которые следует учитывать при планировании и организации дистанционного обучения. Например, *S. Levy*, кроме изменения организационной структуры, считает необходимым учитывать такие факторы, как планирование учебной деятельности, обучение и поддержка персонала и студентов, защита авторских прав и интеллектуальной собственности [3].

Изменение технологий обучения влияет и на развитие понятийного аппарата. Статистика и анализ использования слов и словосочетаний позволяет нам с количественной точки зрения отследить частоту и временные периоды их применения. *M. Baldassarre* [4] указывает, что сервис *Ngram Viewer* позволяет в течение нескольких секунд получить график, описывающий временной ход использования слов.

Так, вопросам изменения применения терминов в системе обучения посвящена статья «*Analyzing the Discourse of Chais Conferences for the Study of Innovation and Learning Technologies via a Data-Driven Approach*», авторы которой на основе сервиса *Ngram Viewer* устанавливают, что в последнее время обучение все больше связано с Интернетом. Исследования проводились в двух языковых корпусах публикаций: на иврите и на английском языке. Согласно получен-

ным результатам, авторы считают, что: «Наиболее частым ключевым словом в рассматриваемом корпусе публикаций было *collaborative learning* (*совместное обучение*) (40 публикаций). За ним последовали *distance learning* (*дистанционное обучение*) и *E-learning* (*электронное обучение*) в 25 и 23 публикациях соответственно» [5].

МЕТОДЫ ИССЛЕДОВАНИЯ

Для полноты исследования эволюции терминов, обозначающих дистанционное, онлайновое и открытое обучение, а также возникновение и развитие его отдельных элементов, нами проведен анализ частоты их упоминания в книжных и справочных изданиях с применением сервиса *Google Books Ngram Viewer*. Инструмент *Ngram Viewer* был специально разработан учеными Гарвардского университета совместно с компанией *Google* для культурологов, лингвистов и других гуманитариев в 2010 г. и сразу же нашел широкое применение [6–8].

В основу нашего исследования легли термины: *дистанционное образование, дистанционное обучение, открытое обучение, открытое образование, дистанционные курсы, онлайн-курсы, онлайновые курсы, онлайн-образование, онлайн-обучение, онлайновое обучение, интернет-образование, интернет-курсы, онлайн-университет, интернет-университет, онлайновые университеты* (всего 15 терминов).

Изучение частоты упоминания терминов проводилось по шести языковым корпусам: русский, немецкий, английский, испанский, итальянский, французский (таблица).

Сопоставимые термины в области дистанционного, открытого и онлайнового образования и обучения для шести языковых книжных корпусов

Группы	Русский	Немецкий	Английский	Испанский	Итальянский	Французский
1. Общие	Дистанционное образование	Fernunterricht	Remote education Distance education	Educación a distancia	Formazione a distanza	Enseignement à distance
	Дистанционное обучение	Fernstudium	Distance learning	Aprendizaje a distancia	Apprendimento a distanza	Enseignement à distance
	Открытое обучение	Training öffnen Offene Ausbildung	Open training	Entrenamiento abierto	Formazione aperta	Entraînement ouvert Formation ouverte
	Открытое образование	Offene Bildung	Open Education	Educación abierta	Open Education	éducation ouverte
2. Специфические	Дистанционные курсы	Fernkurse	Distance Courses	Cursos a distancia	Corsi di distanza	Cours à distance
	Онлайн-курсы	Online-Kurse	Online Courses	Cursos en línea	Corsi online	Cours en ligne
	Онлайновые курсы	Online-Kurse	Online Courses	Cursos en línea	Corsi online	Cours en ligne
	Онлайн-образование	Online-Bildung	Online Education	Educación en línea	Istruzione online	Éducation en ligne
	Онлайн-обучение	Online-Schulung	Online Training	Entrenamiento en línea	Formazione online	Formation en ligne
	Онлайновое обучение	Online-Training	Online training	Capacitación en línea Entrenamiento en línea	Formazione online	Formation en ligne Éducation ouverte
	Интернет-образование	Internet-Bildung	Internet education	Educación en internet	Internet educazione	Éducation à Internet
	Интернет-курсы	Internetkurse	Internet Courses	Cursos de internet	Corsi di Internet	Cours Internet
	Онлайн-университет	Online Universität	Online University	Universidad en línea	Università Online	Université en ligne
	Интернет-университет	Internet Universität	Internet University	Universidad de Internet	Università di Internet	Université Internet
	Онлайновые университеты	Online-Universitäten	Online universities	Universidades en línea	Università online	Universités en ligne

Анализ данных с применением сервиса *Google Books Ngram Viewer* проводился многократно в течение апреля 2016 г. – декабря 2017 г. Необходимо отметить, что формирование языковых публикационных корпусов зависит не только от количества первоисточников, изданных на соответствующем языке. Достаточно часто в языковой публикационный корпус попадает переводная литература, в подсчете участвуют и источники, в которых словосочетания фиксируются в библиографическом списке, при этом в самом издании они раскрываются косвенно, либо отражают аспекты, отличные от исследуемых.

Следует отметить, что сервис Ngram Viewer позволяет одновременно искать не более двенадцати слов. При этом существуют лингвистические особенности применения терминов, например, в русском языке используются различные падежные окончания и для данного сервиса такие слова считаются разными и не группируются в трендовую линию.

Для большей объективности мы проводили исследование с установкой критерия «не использовать признак регистра», что позволяет находить и группировать термины, используемые в начале (с прописной буквы) и в середине предложений (со строчной буквы), как одинаковые. Если в языковом корпусе встречаются термины с разными регистрами, то при формировании графика сервисом Ngram Viewer рядом с этими терминами проставляется обозначение All [6].

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Частотные распределения встречаемости всех исследуемых терминов в русскоязычном корпусе публикаций проведены за период 1920 – 2008 гг. В рассматриваемом масштабе изменения частот встречаемости терминов видимыми оказались 6 из 15 терминов. В дальнейшем, помимо слова «встречаемость» будут также использованы слова «упоминание», «употребление» и «использование».

При анализе рускоязычного корпуса публикаций сервисом *Ngram Viewer* видно значительное преобладание использования терминов «дистанционное обучение» и «дистанционное образование». Доля их употребления настолько высока, что не позволяет отследить тенденции изменений таких терминов, как «дистанционные курсы», «открытое обучение» и «интернет-образование».

Проанализировав все 15 терминов в сервисе *Ngram Viewer* можно однозначно отметить, что чаще употребляемыми являются те, которые носят более общий смысл – такие как *дистанционное обучение*, *дистанционное образование*. Эта тенденция характерна для всех шести языковых корпусов книг.

Из-за значительных вариаций в интерпретации понятийного аппарата и количественных ограничений терминологического запроса в сервисе *Ngram Viewer* для более подробного исследования распределим 15 ранее выделенных терминов на две группы (см. таблицу). Целесообразно начать исследование с более общих терминов, применимых к доступному образованию: *дистанционное образование*, *дистанционное обучение*, *открытое обучение*, *открытое образование*. В этой терминологической группе первые упоминания понятий *открытое обучение* и *открытое образование* относится к середине 1920-х гг.

Необходимо отметить, что в силу специфики русского языка первые упоминания словосочетания *открытое образование* относятся к иным сферам жизни общества. Например, в книге Ю.М. Стеклова «Михаил Александрович Бакунин, его жизнь и деятельность (1814-1876): Раскол в Интернационале» (1927 г.) это словосочетание связано с фрагментом: «... принципов Альянса было сформулировано первое открытое социально-революционное слово, раздавшееся в ... Затем в газете появился ряд его статей, из которых наиболее замечательными являются статьи… об интегральном образовании…» [9, с. 354]. Книга вошла в подборку по словосочетанию *открытое образование* даже несмотря на то, что слова *открытое* и *образование* находятся в разных абзацах одной страницы текста.

В «Большой советской энциклопедии» (том 60, 1934 г.) это словосочетание встречается в следующей трактовке: «Открытое образование общественных союзов и политических группировок, целый ряд правительственных сообщений о революционном движении — все это дезориентировало цензуру и лишало ее твердости и выдержанной ...» [10]. При этом слово *образование* в обоих источниках несет иную смысловую нагрузку, в данном контексте подразумевается *создание* политического объединения.

Применительно к обучению данное выражение появляется в конце 1940-х гг. Например, оно встречается в журнале «Вестник высшей школы» (1947 г.): «Потребность в специальном транспортном образовании в нашей стране возникла в сравнительно далеком прошлом. ... Первоначально это было открытое учебное заведение» [11].

В целом, в 1990-е гг. наблюдался экспоненциальный рост встречаемости в книжном корпусе терминов *дистанционное обучение* и *дистанционное образование*. С 2003-2004 гг. происходит спад популярности этих терминов, что связано, по-видимому, с употреблением других терминов, их заменяющих.

Понятие дистанционного образования в России стало использоваться с середины 1990-х гг. В 1995 г. был издан словарь «Образование взрослых: междисциплинарный словарь терминологии» (В.Г. Онушкин, Е.И. Огарев), в котором дано определение: «Дистанционное образование – международный термин, иногда переводимый как «образование на расстоянии»…» [12]. В этом же году была принята Концепция о создании и развитии единой сети дистанционного образования в России.

С 1993 г., в течение двух лет, Российский независимый университет (сейчас Российский новый университет) издавал журнал «Открытое образование», с 1996 г. Московский экономико-статистический институт начал издавать журнал «Дистанционное образование», который с 2000 г. стал издаваться Российским экономическим университетом имени Г.В. Плеханова под названием «Открытое образование». С 1999 г. Современный гуманитарный университет выпускает журнал «Дистанционное и виртуальное обучение». Все это хорошо коррелирует с ростом встречаемости соответствующих терминов.

Немецкоязычный корпус книг выделяет такие понятия, как: *дистанционное обучение* (*Fernstudium*), *дистанционное образование* (*Fernunterricht*). Значительный пик встречаемости термина «обучение» приходится на 1960–1970 гг. Однако необходимо отметить, что под обучением в научной литературе не всегда понимается получение образования в классическом понимании. В отдельных источниках под открытым обучением понимается спортивное обучение вне помещений.

В корпусе англоязычной литературы эти термины стали применяться немногим позднее, и публикационный пик в исследовании дистанционного обучения приходится на 2002-2005 гг. При этом изучение открытого образования привлекает ученых с конца 1960-х гг. и в 1975 г. на английском языке выходит книга *D. Nyberg* «*The Philosophy of open education*», повлиявшая на появление на графике сервиса *Ngram Viewer* локального максимума. Это фундаментальный труд, посвященный формированию философии понятия *дистанционное образование*. В нем автор говорит, что открытое образование является формой образовательной практики, которая регулируется четырьмя характерными правилами [13]:

«(1) Учащиеся должны участвовать в образовательных мероприятиях по своему выбору;

(2) Учителя должны создать среду, богатую возможностями обучения;

(3) Учителя должны давать индивидуальное обучение ученику на основе того, что ему интересно, но они также должны направлять ученика по установленным планам обучения;

(4) Учителя должны уважать студентов. Выполнение следующих правил означает проявление уважения к ученику …».

Эти правила в дальнейшем, на наш взгляд, легли в основу Болонского процесса. Они фактически проповедовали отказ от фундаментального и целостного образования и переход к фрагментарному образованию, которое в условиях широкого развития сети Интернет привело к формированию у молодежи «клипового» сознания.

В испаноязычном корпусе публикаций понятие *дистанционное образование* появляется в начале 1970-х гг.

Многие испаноязычные книжные издания, зафиксированные в системе сервиса *Ngram Viewer*, хранятся в США. Например, книги «*Los retos de la «educación a distancia»*: *I Seminario-Taller de Teleducación Universitaria de FUPAC*» (1978 г.) [14] и «*Educación*» (1974 г.) [15] находятся в библиотеке Университета штата Пенсильвания, «*La reforma educativa de la Segunda República Española: primer bienio*» (1977 г.) [16] – в библиотеке Калифорнийского университета.

При этом значительное количество книг, в которых упомянуты исследуемые термины, не имеют прямого отношения к системе образования в содержательном понимании. Эти термины фигурируют в библиографических описаниях книг, в аннотациях, в представлении заслуг и этапов обучения авторов, ученых, в публикациях иной направленности и т.п.

Однако имеются и достаточно интересные издания на испанском языке, например, в *Instituto Colombiano para el Fomento de la Educación Superior* в 1986 г. была выпущена книга «*Metodología y estrategias de la educación superior abierta y a distancia: nivel introductorio*» («Методология и стратегия открытого и дистанционного высшего образования») [17].

В италоязычном и франкоязычном корпусах публикаций наблюдается та же закономерность в развитии понятийного аппарата, как и в испаноязычном корпусе. В италоязычном корпусе намного преобладает понятие *дистанционное образование* (*Formazione a distanza*). Незначительный рост частоты встречаемости термина *дистанционное образование* наблюдался с 1980 по 1995 гг., и далее виден сильный рост этой частоты вплоть до 2004 г. В этот период было издано много книг на итальянском языке. По содержанию эти издания относились как к сфере образования, так и к сфере повышения компьютерной грамотности (дистанционное освоение компьютерных программ).

Во франкоязычном корпусе книг термин *дистанционное образование* (*Enseignement à distance*) в начале 1970-х гг. встречался в статьях журналов, докладах, например, в журнале «*L'éducation*» (1971 г.), выпуски 99-105 которого были оцифрованы в 2011 г. (владелец оцифрованного оригинала – Университет штата Пенсильвания).

В испанском, итальянском и французском языковых корпусах книг значительно преобладает термин *дистанционное образование* и по срокам более раннего упоминания, и по частоте встречаемости. В остальных трех языковых группах преобладает термин *дистанционное обучение*. Оба эти термина являются ведущими по частоте встречаемости в первой группе терминов шести рассмотренных языковых корпусов книг.

Развитие понятий в исследуемых языковых корпусах происходит от общего к частному: от понятий открытого и дистанционного образования к более узким и специализированным *дистанционным курсам*, *онлайн-курсам*, что подтверждает ранее выявленную закономерность.

Если не учитывать наиболее распространенные термины первой группы, то динамика развития специфической терминологии в образовании становится более явной. Анализируя статистику упоминания этих терминов, следует отметить, что сервис *Ngram Viewer* показывает в русскоязычном корпусе применение всего лишь двух «специфических» терминов (см. таблицу): *дистанционные курсы* и *интернет-образование*.

Употребление с начала 1990-х гг. исследуемых нами терминов в русскоязычном корпусе обусловлено распространением компьютерной техники и развитием сети Интернет, и по периоду возникновения находится в логической взаимосвязи с понятием *дистанционное обучение*.

В немецкоязычном корпусе частота встречаемости этих терминов немного шире. Сервис фиксирует использование терминов: *дистанционные курсы* (*Fernkurse*), *онлайн-обучение* (*Online-Schulung*), *онлайновое обучение* (*Online-Training*), *онлайн-курсы* (*Online-Kurse*), *интернет-курсы* (*Internetkurse*), причем значительно выделяется понятие *дистанцион-*

ные курсы (*Fernkurse*). Изучив источники, на которые ссылается сервис *Ngram Viewer*, мы установили, что первоначально понятие *Fernkurse* означало *классическое заочное обучение*. Однако в 2000-х гг. достаточно предсказуемо видно использование современных терминов, связанных с возникновением и широким применением компьютерной техники и Интернета. Наибольшую популярность набирают онлайн-курсы, что связано и с развитием системы образования в целом, и с возможностями для человека получить альтернативное образование по конкретному интересующему его направлению или виду деятельности. Отметим, что интернет-курсы – это не только университетские курсы, но и различные мастер-классы и тренинги более широкого спектра.

Динамика частоты встречаемости исследуемых нами терминов, связанна с развитием современных средств коммуникаций и новых технологий, наблюдается и в других языковых корпусах. Например, в англоязычном корпусе книг всё большее внимание уделяется изучению и популяризации онлайн-курсов и онлайн-образования.

Более регулярный и сглаженный вид кривых частотного распределения терминов, формируемых сервисом *Ngram Viewer* в англоязычном корпусе книг, в отличие от всех остальных свидетельствует о том, что данная группа источников в рассматриваемой терминологии очень обширна и, следовательно, репрезентативна.

Отметим уровни насыщения кривых, формируемых сервисом *Ngram Viewer*, для терминов *Online Courses* и *Online Education* в 2008 г., когда были опубликованы работы [18, 19], которые по сути связывают понятие образования с другими направлениями деятельности и развития человека и общества. В этом же году в Гонг Конге была проведена международная конференция «*Hybrid Learning and Education: First International Conference*» по гибридному обучениию, представленному как комбинация традиционного обучения в классе и интернет-обучения в рамках одного метода обучения. Результаты конференции опубликованы [20]. Значительное внимание уделяется и популяризации онлайн образования в массах, например, книга R.L. Du Vivier [21].

Графики частоты встречаемости второй группы исследуемых нами терминов (см. таблицу) в испаноязычном и франкоязычном корпусах книг аналогичны графикам встречаемости этой же группы терминов в немецкоязычном и англоязычном корпусах, однако сервис *Ngram Viewer* по ряду терминов показывает единичные источники. А в корпусе италоязычных книг сервисом фиксируются только термины *онлайн-курсы* и *онлайн-обучение*.

ЗАКЛЮЧЕНИЕ

С помощью сервиса *Google Books Ngram Viewer* нами проведен анализ частоты встречаемости пятнадцати терминов в области дистанционного, открытого и онлайнового образования и обучения для шести языковых корпусов. Все термины для полноты отражения результатов разбиты на две группы: общие и специфические. Отмечена однозначная для всех языковых групп временная характеристика частоты встречаемости исследуемых терминов. Однако не все публикации, учитываемые сервисом *Ngram Viewer*, относятся к рассматриваемой области обучения и образования. Частично они имеют иную смысловую нагрузку, являются омонимами. Наиболее раннему временному интервалу характерно и присуще использование терминов общей группы. Значительно преобладают по частоте встречаемости в шести рассмотренных языковых корпусах два основных термина: *дистанционное образование* и *дистанционное обучение*.

Динамика частоты встречаемости специфических терминов связана с развитием современных средств коммуникации и новых технологий. Максимум их применения приходится на начало 2000-х гг. Трендовые линии всех рассмотренных графиков частотного распределения терминов, формируемых сервисом *Ngram Viewer*, отражают зависимость развития понятийного аппарата от потребностей государства и общества, от проводимой образовательной политики, от социально-экономических условий развития образовательной системы.

Таким образом, наряду с традиционными образовательными процессами всё большее внимание уделяется развитию образования с помощью дистанционных интернет-технологий. Дистанционное образование становится всё более востребованным, обеспечивая общедоступность получения новых знаний.

СПИСОК ЛИТЕРАТУРЫ

1. Keegan D. Foundations of distance education. – London : Psychology Press, 1996.
2. Bates A. W. T. Technology, e-learning and distance education. – Routledge, 2005.
3. Levy S. Six factors to consider when planning online distance learning programs in higher education // Online journal of distance learning administration. – 2003. – Vol. 6, №1. – URL:http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=776E39DD8744D36169E1974622965DF3?doi=10.1.1.495.2749&rep=rep1&type=pdf
4. Baldassarre M. Informazione, Conoscenza, Didattica. La sfida dei big data al mondo della formazione Information, Knowledge, Didactics // The challenge of big data for the world of education. – 2016. – P. 90-112.
5. Silber-Varod V., Eshet-Alkalai Y., Geri N. Analyzing the Discourse of Chais Conferences for the Study of Innovation and Learning Technologies via a Data-Driven Approach // Interdisciplinary Journal of e-Skills and Life Long Learning. – 2016. – Vol. 12. – P. 297-313.
6. Michel J.-B. et al. Quantitative analysis of culture using millions of digitized books // Science. – 2010. – Vol. 331, №1. – P. 176 -182.
7. Jones E. Google Books as a General Research Collection // Library Resources & Technical Services. – 2010. – Vol. 54, № 2. – P. 77 -89.
8. Московкин В.М. Google Books и «культурологические тренды» // Научно-техническая информация. Сер. 1. – 2012. – № 7. – С. 27-34.
9. Стеклов Ю. Михаил Александрович Бакунин, его жизнь и деятельность. 1814-1876: в 3 т., 2-е изд., испр. и доп. – М. : Изд-во Коммунист.

акад., 1926-1927. – 550 с. – Т. 3: Бакунин в Интернационале. (1868-1870 г.).
10. Большая советская энциклопедия. – М.: Изд-во Советская энциклопедия, 1934. – Т. 60. – С. 469
11. Вестник высшей школы: Т. 5. – М.: Изд-во Советская наука, 1947.–
12. Онушкин В.Г., Огарев Е.И. Образование взрослых: междисциплинарный словарь терминологии. – М.: Изд-во Российская академия образования, Институт образования взрослых, 1995. – 231 с.
13. Nyberg D. The Philosophy of open education. – London; Boston: Routledge & K. Paul, 1975. – 213 p.
14. Arrien J.B. Los retos de la «educación a distancia» // I Seminario-Taller de Teleducación Universitaria de FUPAC, V Seminario Latinoamericano de Teleducación Universitaria. – Federación de Universidades de América Central y Panamá (FUPAC), 1978. – P. 264.
15. Educación. – Departamento de Asuntos Educativos, Secretaria General de la Organización de Estados Americanos, 1974. – URL: https://books.google.ru/books?id=sMxXAAAAYAAJ&q=%22educaci%C3%B3n+abierta%22&dq=%22educaci%C3%B3n+abierta%22&hl=ru&sa=X&ved=0ahUKEwiQ6tq33q7aAhWNxKYKHdqUAEEQ6AEIKDAA
16. Pintado A.M. La reforma educativa de la Segunda República Española: primer bienio. – Santillana, 1977. – Vol. 15.
17. Instituto Colombiano para el Fomento de la Educación Superior. Metodología y estrategias de la educación superior abierta y a distancia: nivel introductorio. – Universidad Abierta ya Distancia, 1986.
18. The theory and practice of online learning / ed. T. Anderson. – Athabasca University Press, 2008.
19. Economics of distance and online learning: Theory, practice and research / eds. W. J. Bramble, S. Panda. – Routledge, 2008.
20. Hybrid Learning and Education: First International Conference, ICHL 2008 Hong Kong, China, August 13-15, 2008 Proceedings / eds. J. Fong, R. Kwan, F.L. Wang. – Berlin: Springer-Verlag Heidelberg, 2008.
21. DuVivier R.L. 100% Online Student Success. – Cengage Learning, 2008.

Материал поступил в редакцию 20.12.17.

Сведения об авторах

МОСКОВКИН Владимир Михайлович – доктор географических наук, ведущий эксперт Центра стратегического развития и наукометрических исследований, профессор кафедры мировой экономики Белгородского государственного национального исследовательского университета
e-mail: moskovkin@bsu.edu.ru

САПРЫКИНА Татьяна Валерьевна – кандидат экономических наук, доцент, доцент кафедры финансов, инвестиций и инноваций Белгородского государственного национального исследовательского университета
e-mail: saprykina@bsu.edu.ru

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.65:54.021

С.В. Трепалин, Ю.Е. Бессонов, Б.С. Фельдман, Н.И. Чуракова

Использование химических идентификаторов InChI и InChIKey для поиска химических структур в базах данных

Рассматриваются вопросы построения международных химических идентификаторов IUPAC InChI и InChIKey и использования их для поиска информации о химических соединениях в базах структурных данных по химии. Приведено обоснование использования указанных идентификаторов в технологии обработки структурной химической информации ВИНИТИ РАН.

Ключевые слова: *InChI, InChIKey, химическая информатика, базы структурных данных по химии*

ВВЕДЕНИЕ

Современные базы данных химических соединений и материалов немыслимы без поиска по химическим структурам. Поиск решает две основные задачи. Во-первых, он предоставляет химику-исследователю возможность пользоваться массивами доступных в *on-line* режиме баз данных по химии, а, во-вторых, облегчает задачу пополнения своих баз данных из внешних источников. В последнем случае поиск помогает установить, имеется ли какое-либо соединение в существующей базе данных и, если имеется, то сделать ссылку на существующую структуру. При отсутствии соединений в базе данных они добавляются, при этом осуществляется проверка на внутренние дубликаты среди добавляемых соединений. Ключевую роль в поиске химических соединений играют методы их идентификации.

ПОСТАНОВКА ЗАДАЧИ

Химическую структуру можно представить как группу объединенных связями атомов. Те из них, которые уверенно и однозначно обрабатываются существующими в мире программами, должны подчиняться теории валентных схем. Согласно этой теории каждый атом имеет свойство валентности. Валентность – число связей, которое образует атом. Например, водород образует одну связь, а кислород две. Соответственно, валентность водорода равна единице, а кислорода – двум. Валентность может быть переменной. Например, фосфор может быть 3-валентным и 5-валентным.

Соединения, которые подчиняются правилу валентных схем, могут быть представлены в виде графа, где каждая вершина соответствует атому, а каждое ребро – связи. Пара атомов может быть связана только одним ребром. В том случае если несколько валентностей участвует в образовании связи, говорят о порядке связи – двойная, тройная.

Таким образом, химическую структуру, подчиняющуюся методу валентных схем, можно представить как окрашенный граф, где цвет вершины соответствует положению атома в периодической таблице, а цвет ребра – порядку связи. Все математические операции с графами (работа с циклами, поиск путей, изоморфное вложение) применимы и к простейшим химическим структурам.

Химия вносит свои коррективы в эту «идеальную» модель. В структуре химического соединения могут присутствовать атомы с теми или иными атрибутами (определенный заряд, свободный радикал, тот или иной изотоп элемента). Далее, наличие свойства ароматичности стирает границу между чередующимися двойными и одинарными связями, делая их как бы полуторными. Наиболее часто встречающимся свойством химических структур, усложняющим использование теории графов, является таутомерия, когда для одной или нескольких вершин нельзя сказать, с какими конкретными вершинами они связаны. Чаще всего таутомерия связана с подвижным атомом водорода, но в ряде случаев наблюдается более сложная таутомерия, как, например, в сахарах.

Иногда химического графа недостаточно для описания химической структуры, которая подчиняется методу валентных схем. Это происходит потому, что наше пространство трехмерно и атомы в химических структурах имеют *X, Y* и *Z* координаты. Химический граф вообще не имеет отношения к координатам атомов и указывает только на то, какой атом с каким связан. Ради удобства химический граф изображают

на плоскости. Именно в этом случае и возникают проблемы – когда формально идентичные химические графы соответствуют структурам с различающимися трехмерными координатами атомов. В таких ситуациях говорят об изомерах. Наиболее распространенные пространственные изомеры – стереоизомеры, цис- и трансизомеры, возникающие при заторможенном вращении вокруг двойной связи или в циклических соединениях. Описание подобных структур в виде химического графа потребовало добавление специальных стереосвязей – связь вверх, связь вниз, неопределенная связь, двойная неопределенная. Порой стереоизомерию невозможно представить на химическом графе, например, в случаях альфа-, альфа'-бинафтилов. Кроме того, имеется ряд соединений, которые не подчиняются теории валентных схем. Это значит, что невозможно представить структуру в виде окрашенного графа. К таким соединениям относятся ионные кристаллы, сплавы, бораны, многие комплексы металлов. Сравнение структур таких соединений плохо поддается алгоритмизации и требует много ручной работы.

И, наконец, в промышленности имеют дело не с химическими структурами, а с материалами. Разные материалы могут иметь одинаковый состав, например, графит и алмаз. Многие материалы представляют собой композиты и смеси – цемент, целлюлоза, пальмовое масло. Определенные проблемы имеются также в описании полимеров.

Следует особо отметить, что возможность наблюдения различных изомеров химического соединения зависит от экспериментальных методов, использованных для установления его структуры. Чем меньше время наблюдения в эксперименте, тем больше в химических структурах наблюдается разнообразия.

К быстрым методам относятся, например, ИК-, УФ-спектроскопия. Время взаимодействия фотонов с молекулами порядка пикосекунд, и поэтому в структурах фиксируются особенности, которые невозможно представить в методе валентных схем. Это могут быть возбужденные электронные состояния молекул, которые ответственны за оптические переходы в спектрах. Геометрия небольших молекул в возбужденных электронных состояниях заметно отличается от основного. Отражением существования "быстрых" структур являются, например, полосы поглощения в ИК-спектрах транс- и гош- изомеров 1,2-дихлорэтана. Здесь изомерия возникает в результате заторможенного вращения вокруг одинарной С-С связи. Данный вид изомерии не наблюдается в более медленных методах. В частности, в ЯМР-спектроскопии, где время спиновой релаксации составляет несколько секунд. Однако в ЯМР-спектрах фиксируются различные изомеры амидов карбоновых кислот, возникающие за счет заторможенного вращения вокруг C-N одинарной связи. Разделить изомеры амидов, как правило, не удается из-за быстрого перехода между изомерами и достижения равновесия. И, наконец, если соединение имеет долгоживущие изомеры, которые можно выделять и исследовать отдельно, то при создании химических баз данных их необходимо обрабатывать как индивидуальные изомеры.

Описание структур, различающихся в результате рассмотрения быстрых процессов, представляет большой интерес в научно-исследовательской работе. Однако оно не является необходимым для баз данных, ориентированных на синтетическую органическую химию. Такие базы являются наиболее востребованными, поскольку они содержат информацию как об исходных соединениях для синтеза, так и о синтезированных продуктах, предназначенных для испытания определенных активностей или о других веществах, используемых в самых различных областях промышленности. Поэтому требуется механизм описания долгоживущих структур, которые можно выделить и хранить продолжительное время. Если две разные структуры при небольших сроках хранения превращаются в одну и ту же структуру (или смесь) – то такие химические структуры (или изомеры) следует рассматривать как одинаковые.

Эти сложности приводят к тому, что в общем виде работа с химическими структурами представляет собой нетривиальную задачу, для решения которой необходимо значительное количество ручного труда при идентификации одинаковых структур с учетом наличия изомеров. Особенно это становится актуальным для больших баз данных, содержащих миллионы структур.

ПОИСК ПО ТОЧНОЙ ХИМИЧЕСКОЙ СТРУКТУРЕ

Пополнение больших баз начинается с установки факта – имеется данная структура в базе или нет. Для структуры, которая представлена в виде графа, после идентификации ароматических связей можно выполнить сравнение на идентичность графов [1]. Однако операции на графах требуют значительного количества вычислительных ресурсов, которое с ростом сложности задач возрастает экспоненциально. Поэтому альтернативой сравнению графов на совпадение является использование линейных кодов химических структур.

Линейный код химической структуры представляет собой строку, которая после преобразований по заданным правилам дает матрицу связности – набор атомов и связей между ними. В частности, название структуры является линейным кодом. Но названия не подходят для сравнения структур на идентичность, поскольку основные номенклатуры, как *IUPAC* [2, 3], так и *CAS* [4] допускают неоднозначную генерацию названий. Это усугубляется тем, что химики часто пользуются смесью номенклатур, а также тривиальными названиями. Кроме того название ориентировано на понимание этой информации человеком – воссоздание матрицы связности из названия в общем виде не решено однозначно [5]. Следовательно, для решения проблемы сравнения двух структур необходимо ориентироваться на машинно-читаемые названия – линейные кодировки.

Первым представлением структуры в виде строки следует считать линейную нотацию Висвессера [6] (*WLN*). Широкое распространение получила нотация *SMILES* (*Simplified Molecular Input Line Entry System*) [7]. Линейная кодировка *SYBYL* [8] позволяет описывать не только химические структуры, но и реакции.

Имеются и другие линейные нотации химических структур, например, *MCDL* [9], *SEFLIN* [10].

В начале 2000-х гг. *IUPAC* запустил проект по созданию линейной нотации *InChI* [11]. В *InChI* учтены неудачи и проблемы предыдущих линейных нотаций, которые делали невозможным или затруднительным сравнение химических структур посредством сравнения линейных нотаций. Кроме того в *InChI* используется обширный набор правил для нормализации химических структур, что позволяет генерировать одинаковую линейную нотацию для одинаковых структур, но задаваемых разными способами. Кроме того, *InChI* имеет ряд других преимуществ. В частности, для сравнения структур на совпадение можно сравнивать не *InChI*-строки, которые имеют варьируемую и достаточно большую длину, а хэш-строку фиксированного размера 27 байт – *InChIKey* [12].

ЛИНЕЙНАЯ НОТАЦИЯ InChi

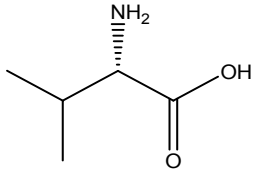
Для генерации линейной нотации *InChI* требуется химическая структура, задаваемая в простом, легко реализуемом формате. Программы для генерации – двоичные библиотеки **.dll* (32 и 64 разрядная *Windows*), **.so* (32 и 64 разрядный *Linux*), графический интерфейс для *Windows* – *winchi-1.exe*, а также исходные коды на языке C++ и детальная документация доступны на официальном сайте группы *InChI Trust* [13]. Последняя версия на момент написания данной статьи 1.05 от 27 января 2017 г. использовалась для тестов, описанных в настоящей статье.

Коды и двоичные файлы распространяются свободно и могут быть использованы в любых приложениях, в том числе коммерческих. Для работы с *InChI* не требуется запрашивать разрешения у владельца группы *InChI Trust* [14]. При этом *InChI* и *InChIKey* являются торговыми марками и их можно использовать только для строк, которые генерируются оригинальными программами InChI Trust. Внесение модификаций в исходные коды приводит к генерации линейных нотаций, отличных от оригинальных. При этом теряется право на использование зарегистрированных знаков *InChI* и *InChIKey*. Таким образом *InChI Trust* пытается обезопасить себя от ситуации, которая случилось с линейной нотацией *SMILES*. В результате внесения бесконтрольных изменений в форматы *SMILES* появилось, как минимум, два широко распространенных формата – *Daylight* [15] и *Open Eye* [16], а также ряд других программных продуктов. *SMILES*-строки, генерируемые различными программными продуктами, различаются. Кроме того, обнаружены проблемы с восстановлением матрицы связности для строк, сгенерированных в другом программном продукте.

Линейная кодировка *InChI* состоит из слоев, разделенных символом "/" (слэш). Пример *InChI* строки для *L*-валина приведен на рис. 1.

Стереохимия представляется несколькими слоями и завершается парой " m0/s1" – признак окончания стереохимической информации; "t4-" – стереохимическая информация для тетраэдрического атома. Могут быть слои для стереохимии при двойной связи. Также определены слой "/i", который содержит информацию об изотопном составе и слой "/f", содержащий информацию о таутомерах – позиции присоединения протона; слой "/q" с последующим числом – общий заряд молекулы. Подробно эти слои описаны в [17].

Правила заполнения слоев и генерации *InChI*-строки не имеют существенного значения, так как это осуществляется программами, поставляемыми *IUPAC*. Восстановление матрицы связности из строки *InChI* также осуществляется программой *IUPAC*. Но поскольку строка *InChI* не содержит координат атомов, их необходимо восстанавливать при помощи внешних программ [18, 19] . При этом часто возникают проблемы, особенно в случае полициклических соединений. По этой причине базы данных хранят не в виде *InChI*-строк, а в форматах, предусматривающих сохранение координат атомов. Однако пользователям баз данных с открытым доступом в Интернете строку *InChI* следует предоставлять. Современные программы, например, [20], понимают данные *InChI*, помещенные в буфер обмена. Используя операцию копировать/вставить можно легко перенести химическую структуру. Если же разрабатывается частная база, владельцы которой не хотят демонстрировать химические структуры в электронно-читаемых форматах неограниченному кругу лиц, то *InChI*-строки (а также другие линейные нотации) не следует предоставлять пользователям.



InChI=1S/C5H11NO2/c1-3(2)4(6)5(7)8/h3-4H,6H2,1-2H3,(H,7,8)/t4-/m0/s1

Рис. 1. Строка *InChI* для L-валина

где: первый слой "*1S*" – версия *InChI*

второй слой "*C5H11NO2*" – молекулярная формула

третий слой "*c1-3(2)4(6)5(7)8*" – кодировка матрицы связности

четвертый слой "*h3-4H,6H2,1-2H3,(H,7,8)*" – кодировка числа атомов водорода, присоединенных к атомам. Допускается кодировка подвижных протонов, которые относятся к нескольким центрам.

Строка *InChI* может быть сгенерирована с различным набором опций, например, учитывать таутомеры или нет, удалять или не удалять связи с металлом и др. Детальное описание этих опций можно найти в документации [21]. При включении или отключении какой-либо опции получаются разные строки *InChI*. Для сравнения содержимого разных баз данных и поиска ресурсов в Интернете был определен фиксированный набор параметров. Строка *InChI*, которая генерируется с использованием этих параметров, называется стандартной (*standard InChI*); без них – нестандартной (*non-standard InChI*). Набор параметров включает в себя кодировку скелета, стереоинформацию с учетом только абсолютной стереоконфигурации (или отсутствие стереоинформации), учет таутомерии, зарядов, изотопов и делокализованных протонов. Связи с атомами металла удаляются перед генерацией стандартной *InChI*. Этим критериям удовлетворяет большинство химических баз данных. Практически все ресурсы в Интернете (где используется *InChI*) созданы с использованием стандартной строки *InChI*. К сожалению, в список параметров стандартной строки InChI не попала кето-енольная таутомерия.

В линейной кодировке *InChI* записаны номера атомов, которые являются каноническими, то есть одинаковыми для разных исходных представлений одной и той же химической структуры. Для канонизации номеров атомов должна быть не только предварительно нормализована запись структуры, но и учтены ее стереохимические особенности.

НОРМАЛИЗАЦИЯ ХИМИЧЕСКИХ СТРУКТУР

Перед генерацией линейной нотации химическую структуру необходимо привести к стандартному виду. Одна и та же химическая структура может быть представлена, как правило, многими способами. Хороший пример – ацетат натрия. Это соединение часто изображают с ковалентной, а часто с ионной связью (рис. 2). Процедура приведения химических структур к нормальному виду называется нормализацией.

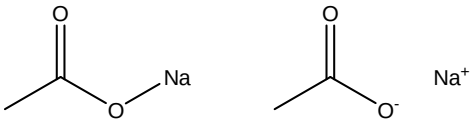


Рис. 2. Различные изображения структуры ацетата натрия

InChI имеет, пожалуй, самую мощную схему нормализации структур среди созданных к настоящему времени приложений по обработке химических структур. Нормализация *InChI* приведена в таблице с разрешения авторов [17].

Сложными случаями являются 1,5- и 1,7-протонные сдвиги, но они встречаются достаточно редко. Возникающая за счет 1,7-миграции протона нитрозо-оксимная таутомерия показана на рис. 3.

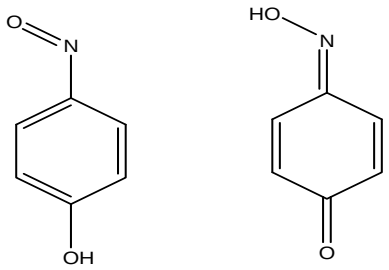
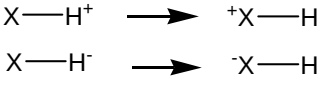
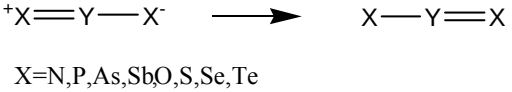
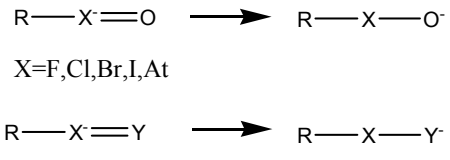
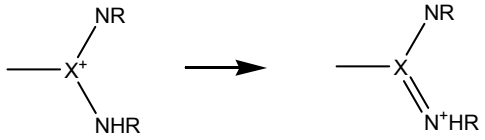
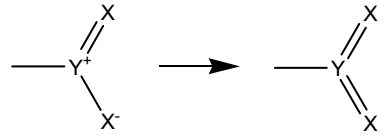
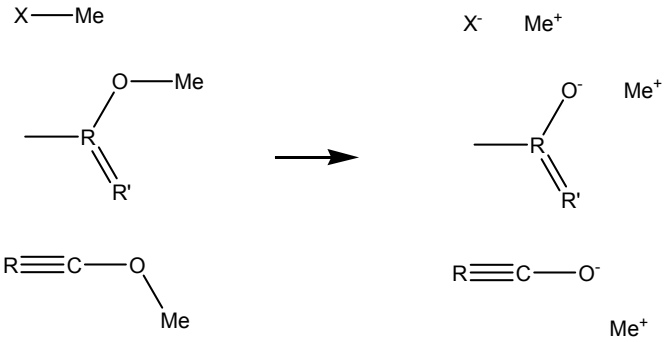
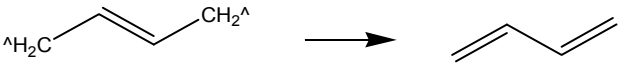
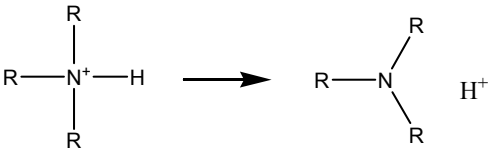
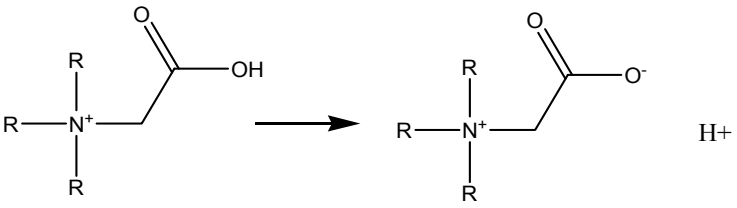
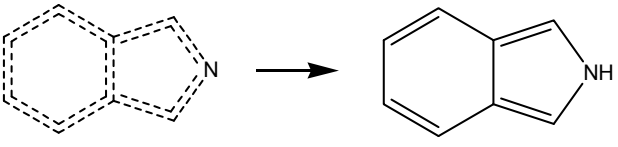


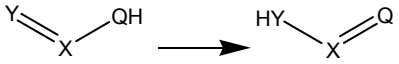
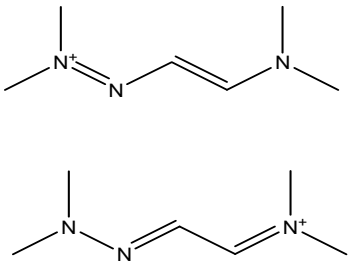
Рис. 3. Пример нитрозо-оксимной таутомерии

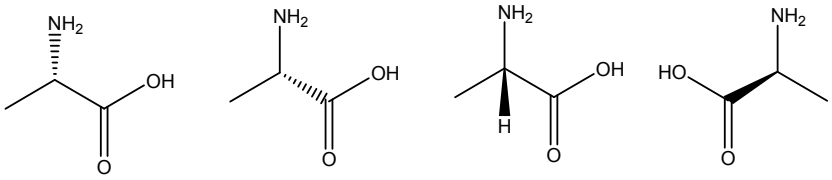
Случаи 1,5- и 1,7-протонных сдвигов достаточно редки. Кроме того, в настоящее время проект *InChI* имеет рабочую группу, которая занимается учетом таутомеров с дальней миграцией протонов [22]. А вот с кето-енольной таутомерией ситуация сложнее из-за гораздо большей распространенности, что приводит к проблемам, которые будут проанализированы ниже.

Схема нормализации InChI

Схема	Описание
	Перемещения заряда с водорода на тяжелый атом
 X=N,P,As,Sb,O,S,Se,Te	Конверсия фрагмента с разделенными зарядами к нейтральному
 X=O,S,Se,Te Y=S,Se,Te	Уменьшение числа заряженных атомов посредством увеличения валентности

 X=F,Cl,Br,I,At X=S,Y=O; X=Se,Y=S,O; X=Te, Y=O,S,Se	Перемещение отрицательного заряда от галогена к кислороду в оксоанионах
 X=C,S	Увеличение порядка связи и перемещение положительного заряда для генерации иминного азота
 Y=N,P,As,Sb; X=O,S,Se,Te	Уничтожение противоположных зарядов с увеличением валентности
	Все связи с атомом металла убираются
	Элиминирование радикалов
	Удаление протона с заряженного гетероатома. Схема упрощенная – полное описание в [17]
	Удаление протона с нейтрального гетероатома, если общий заряд молекулы положительный
	Конверсия ароматических связей к чередованию одинарных и двойных (Кекулизация)

	Простейший учет таутомеров, возникающих за счет 1,3-сдвигов протона. В стандартной версии не учитывает кето-енольной таутомерии
	Детектирование положительного заряда, который может мигрировать



InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1

Рис. 4. Различные изображения L-аланина.

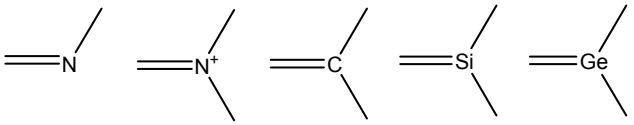


Рис. 5. Двойные связи, которые образуют стереоизомеры.

В текущем релизе *InChI* имеются опции *KET и 15T* для учета кето-енольной и 1,5-миграции протона таутомерии [23]. Однако, строки *InChI*, сгенерированные с использованием этих опций, являются нестандартными, и их в общем виде невозможно сравнивать с обычными стандартными идентификаторами. В обозримом будущем ожидается выход *InChI* версии 2, где по умолчанию будет включен анализ кето-енольной таутомерии[1].

СТЕРЕОХИМИЧЕСКАЯ ИНФОРМАЦИЯ

Строка *InChI* при наличии стереоцентров содержит описание стереохимической информации. Эта информация учитывается при генерации канонической нумерации атомов и является уникальной для каждого стереоизомера независимо от того, как стереоинформация изображается в исходных структурах. Уникальная строка *InChI* химического соединения L-аланина, имеющего 4 различных стереоизображения, приведена на рис. 4.

Программное обеспечение, поставляемое *InChI Trust*, обрабатывает стереоцентры для тетраэдрических атомов, двойных связей и алленов. Стереохимическая информация извлекается как из традиционных связей (*Up, Down, Either*), из изображения заместителей при двойной связи, так и из 3D координат атомов. Помимо традиционного использования связей *Up* и *Down*, когда тонкий конец связи указывает на атом, имеющий стереоконфигурацию, применяется перспективное изображение – при котором оба конца *Up*- или *Down*-связи указывают стереоцентры. Это достигается путем использования переключателя при генерации строки *InChI*.

Список двойных связей, которые образуют стереоизомеры, приведен на рис. 5.

[1] Частное сообщение профессора МГУ им. М.В. Ломоносова (Химический факультет, Кафедра аналитической химии) И.В. Плетнева.

Атомы, которые могут приводить к появлению стереоизомеров с тетраэдрической конфигурацией:

Четырехкоординационные C, Si, Ge, N+, P+, As+, B – , Sn, N, P, S, S+, Se, Se+

Трехкоординационные S, S+, Se, Se+ и N в трехчленном цикле.

Стереоизомеры в кумуленах учитываются только для цепочек из атомов C, Si, Ge и смешанных цепочек из этих атомов. Другие типы стереоизомеров, например, пространственно затрудненные бинафтилы, плоские квадратные комплексы не поддерживаются.

КАНОНИЧЕСКАЯ НУМЕРАЦИЯ АТОМОВ В СТРУКТУРЕ

Химическая структура может быть изображена десятками и сотнями способов, при этом нумерация атомов произвольна. Каноническая нумерация подразумевает такие преобразования, которые приводят к идентичной нумерации атомов независимо от того, какая нумерация была в исходной структуре. Разные алгоритмы приведения структуры к канонической дают различные нумерации, но в пределах одного и того же алгоритма нумерация атомов в структуре всегда идентична независимо от того, в каком порядке определены атомы и связи в исходных структурах.

InChI использует алгоритм Маккея [24], публикация [17] также содержит детальное описание алгоритма генерации канонической нумерации атомов уже с химической точки зрения. Этот алгоритм распространяется в свободном доступе, в отличие от *SMILES*, где используется приватный алгоритм *SMILES2* [25]. Это одна из главных причин проигрыша линейной нотации *SMILES* линейной нотации *InChI*.

InChIKey – ХЕШИРОВАННАЯ ЗАПИСЬ СТРОКИ InChI

Как было отмечено выше, строка *InChI* является уникальной для каждой химической структуры, при этом одинаковой для идентичных, но по-разному изображенных структур и таутомеров. Поэтому, сравнивая строки *InChI,* можно установить, имеется уже структура в базе или нет. Главный недостаток процедуры сравнения заключается в том, что строки *InChI* имеют переменную длину и могут быть очень большими – несколько тысяч символов длиной. Манипулировать такими строками достаточно сложно. Объясняется это длительными по времени операциями загрузки, сравнения и др.

Выход был найден. Вместо полной *InChI* строки можно использовать хэшированные значения. Такая технология получила название *InChIKey* [12]. *InChIKey* представляет собой строку длиной 27 символов. В ней разрешаются только заглавные английские буквы и разделители – знак "минус". Формат *InChIKey* следующий [17]:

AAAAAAAAAAAAAA-BBBBBBBBFV-P

Здесь:

AAAAAAAAAAAAAA – кодировка скелета. Используется *SHA-2 (Secure Hash Algorithm*) алгоритм, который возвращает 256-битовый хэш-код [26]. Для кодировки скелета берутся первых 65-бит. Далее осуществляется base-26 перекодировка двоичных значений и, как результат, получаем строку длиной 14 символов.

BBBBBBBB – кодирование стереоконфигурации, изотопов, точных позиций мигрирующих протонов и другое [17]. Блоки *InChI*-строки, не относящиеся к скелету, хэшируется *SHA-2*. Берутся первые 37 бит и при помощи перекодировки *base-26* получается строка длиной 8 символов.

F – принимает два значения: S, если исходная строка *InChI* генерировалась со стандартным списком параметров или N, если использовалась нестандартная строка *InChI*.

V – номер версии – сейчас везде A (Версия 1).

P – признак катиона или аниона. N – структура электронейтральна, M: -1, L: -2, K: -3, ... , B: -12, O: +1, P: +2, ... , Z: +12. Значению A соответствует заряд либо больше +12, либо меньше -12.

Полные хэш-коды длиной 256 бит обрезаются для того, чтобы получить компактные кодировки химических структур. Компактность важна, поскольку для быстрого поиска по химической структуре предполагается загрузка и хранение *InChIKey* в ОЗУ.

Для сравнения двух химических структур без учета стереоконфигурации, изотопного состава и общего заряда молекулы достаточно сравнить первые 14 символов *InChIKey*. Если необходимо учесть стереоконфигурацию и изотопный состав, то сравниваются как первые 14 символов хэш-кода скелета, так и 8 символов хэш-кода стереоинформации – отрывать хэш-код специальных свойств молекул от хэш-кода скелета нельзя! Если необходимо знать точное соответствие структур, сравниваются целиком *InChIKey* строки.

Сравнение строк *InChIKey* методом перебора – нелогичное решение. Как к любой строке, к *InChIKey* применимо понятие метрики: можно сказать какая из двух строк больше, какая меньше, или определить, что строки равны. Поэтому сначала все строки *InChIKey* базы данных загружаются в ОЗУ. Чтобы оперировать с данными в ОЗУ на *SQL*-ориентированных приложениях, чаще всего бывает достаточно объявить поле индексным. Далее осуществляется сортировка строк – формируется вспомогательный массив, содержащий ссылки на строки в алфавитном порядке. Поиск в такой системе осуществляется при помощи алгоритма бисекций [27], скорость сходимости которого пропорциональна $log_2(N)$, где N – размер базы данных.

Следует сделать два замечания по поводу реализации поиска для больших баз данных.

1. Сравнение целых чисел (32-битовой переменной) осуществляется за то же время, что и сравнение байта (8-битовой) переменной, поэтому для ускорения времени сравнения надо сравнивать не 27 байт *InChIKey*, а семь переменных длиной 4 байта каждая. В разных языках программирования это достигается разными путями. В любом случае для хранения *InChIKey* надо использовать 28 байт ОЗУ, дополняя последний байт нулем. Модельные расчеты показали, что скорость сравнения увеличивается примерно в 2,5 раза. Этот фактор важен для больших баз данных, которые изменяются без остановки работы приложения и обслуживания запросов пользователей. После каждого дополнения или изменения требуется пере-

сортировка базы для работы алгоритма бисекций и выигрыш в 2,5 раза становится заметным.

2. Хранение *InChIKey* в ОЗУ потребляет много ресурсов – для базы из 100M записей потребуется 2.7G ОЗУ. Это заметная потеря ресурсов. С другой стороны, если хранить *InChIKey* в виде оригинальных двоичных значений (65 бит скелет+37 бит специальная информация), то для хранения одной строки потребуется 102 бита или 13 байт. Для хранения базы 100M записей потребуется 1.3G ОЗУ – экономия более чем в два раза. При этом в два раза увеличивается скорость сортировки и поиска за счет уменьшения числа сравнений и в два раза возрастет скорость загрузки базы в ОЗУ с носителей. Двоичное представление *InChIKey* присутствует в исходных кодах, поставляемых *IUPAC*, и его можно извлечь. Но это уже будет вмешательство в программу. К тому же такое вмешательство придется осуществлять каждый раз после выхода новой версии или обновления *InChI*. Было бы разумно, чтобы *IUPAC* добавил в интерфейс метод для получения двоичного представления *InChIKey*.

СОВПАДЕНИЕ InChIKey ДЛЯ РАЗЛИЧНЫХ СТРУКТУР

InChIKey представляет собой хэш от *InChI* строки и как для любого хэш-кода имеется ненулевая вероятность генерации одинаковых *InChIKey* для разных строк *InChI*. Эта проблема в англоязычной литературе получила название "*collision*" – коллизия, (синонимы: конфликт, столкновение). В самом деле, хэш-код скелета имеет размер 65 бит, что дает динамический диапазон $2^{65}\approx3{,}7*10^{19}$ допустимых значений. Между тем, оценочное число топологических изомеров для алкана тетрапентаконтана ($C_{54}H_{110}$) равно $5{,}79*10^{19}$, что превышает динамический диапазон топологического блока *InChIKey* [28]. То есть тетрапентаконтан гарантированно имеет множество пар топологически отличающихся стереоизомеров с одинаковыми *InChIKey*.

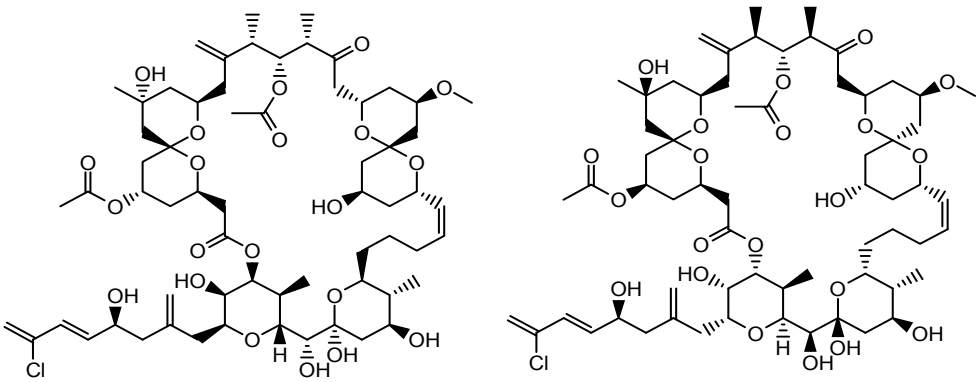
На самом деле для достижения первых коллизий совсем не обязательно генерировать базу данных с записями в количестве 10^{20}. В работе [12] совпадения *InChIKey* анализируются теоретически и экспериментально. Основные результаты и выводы этого исследования:

- на модельных базах данных размером около 100M записей выполнен расчет *InChIKey* и показано, что значения *InChIKey* равномерно распределены по всему динамическому диапазону;
- вероятность совпадения первого блока *InChIKey*, описывающего топологию, для базы данных случайных соединений с числом записей 1000M равна 1,3%;
- вероятность совпадения первого блока *InChIKey* для базы с числом записей 100M равна 0,014%.

Столь большая разница динамического диапазона (10^{20}) и размера базы (10^{10}), где уже реально можно наблюдать одинаковые значения *InChIKey* для разных соединений объясняется статистикой, которая называется «парадокс дня рождения». Задача формулируется следующим образом: собралась группа людей, и какова вероятность события, что хотя бы у двух человек совпадут даты рождения, независимо от года рождения. Очевидно, что если группа из одного человека, то вероятность такого события равна нулю. Если группа из 366 человек (високосные года не учитываются), то вероятность такого события равна единице. Для того чтобы вероятность была 0,5, на первый взгляд, требуется группа из 183 человек. На самом деле для того чтобы в группе людей у двух человек совпали дни рождения с 50% вероятностью, достаточно собрать группу из 23 человек [29].

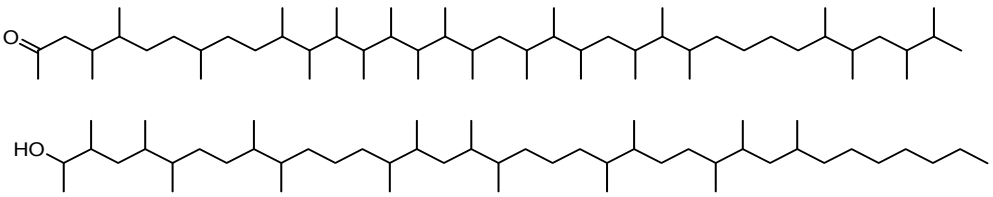
На примере спонгистатина (рис. 6) была выполнена генерация и расчет *InChIKey* различных стереоизомеров для проверки совпадения значений второго блока *InChIKey* для различных стереоизомеров. Ранее в отношении спонгистатина было обнаружено совпадение вторых блоков [30]:

Для этого блока длиной 37 бит 50% вероятность совпадения следует ожидать для выборки размером $3{,}7*10^{5}$. Спогистатин имеет 26 хиральных центров (включая двойные связи) и может иметь $6{,}7*10^{7}$ изомеров, что заметно больше 50% порога. Генерировались различные стереоизомеры, рассчитывались *InChIKey*, и изучалась частота совпадений второго блока *InChIKey*. Теоретическая и наблюдаемая частота коллизий совпадают с учетом равномерного распределения значений второго блока.



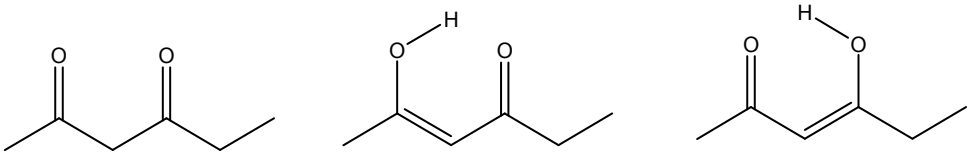
ICXJVZHDZFXYQC-RAZYNMGUSA-N

Рис. 6. Стереоизомеры спонгистатина



OCPAUTFLLNMYSX-UHFFFAOYSA-N

Рис. 7. Различные структуры с одинаковым первым блоком *InChIKey*



NDOGLIPWGGRQCO-UHFFFAOYSA-N

VGWFDUZMKAWCNX-PLNGDYQASA-N

HZEJQMZFNPCDQP-XQRVVYSFSA-N

Рис. 8. Три основные структуры поликетонов.

В настоящее время найден единственный пример совпадения первого блока *InChIKey*, когда разные химические структуры (рис. 7), имеющие разные строки *InChI* дают одинаковый *InChIKey* [31].

Этот факт подтверждает следующие теоретические выводы: совпадения первого блока *InChIKey* должны начинаться с баз данных размером 10^9 записей; распределение генерируемых *InChIKey* по динамическому диапазону равномерно. На наш взгляд равномерное распределение – это замечательная особенность, которой удалось добиться авторам *InChIKey* технологии, и которая сделала ее применимой при относительно небольшом динамическом диапазоне *InChIKey*.

ПРОБЛЕМЫ КЕТО-ЕНОЛЬНОЙ ТАУТОМЕРИИ

Как уже упоминалось ранее, в набор параметров для генерации стандартной строки *InChI* не включен учет кето-енольной таутомерии. Это приводит к экспоненциальному росту числа структур и различных *InChIKey* для поликетонов. На рис. 8 показаны три основные структуры (следует ожидать содержание каждой из них >1% в равновесной смеси при комнатной температуре для теплоты енолизации – 13kJ/mol [32]).

Это одна и та же структура в различных таутомерных формах. Помимо этих химических структур можно предложить еще 2 высокоэнергетические структуры енолов и 3 структуры биенолов, различающиеся топологией. С учетом стереохимии вращения вокруг двойной связи C=C-OH (Z/E/не определено) получим 46 возможных изомеров (включая основные, стерео и высокоэнергетические изомеры) для 2,4-диоксогексана.

Все эти изомеры будут иметь отличающиеся *InChIKey*. Следовательно, все они попадут в регистрационную систему или, например, будут рассматриваться как различные структуры для скрининга. Это приведет как к увеличению средств на приобретение формально разных, но реально расфасованных из одной банки соединений, так и к росту затрат на скрининг пропорционально количеству фиктивных изомеров.

С увеличением числа кето-групп (трикетоны, и т.д.) число изомеров возрастает экспоненциально. Следовательно, необходима фильтрация такого типа химических структур. Фильтрация может быть достигнута включением опции *KET* при генерации *InChI* и *InChIKey*. Однако полученные строки *InChI* и *InChIKey* не являются стандартными и их нельзя использовать для связи с химическими ресурсами Интернета.

Для решения этой проблемы перед работой со структурами предлагается их предобработка. А именно – перед занесением структуры в базу (поиска в базе) она нормализируется. Нормализация заключается в том, что если находится гидроксильная группа при двойной неароматической связи, то она конвертируется в кето-группу. Такая конвертация однозначна и просто выполнима. А далее осуществляется расчет стандартной строки *InChI* и *InChIKey*. При таком подходе енольные формы не попадут в базу данных и их не будет в запросах. И, поскольку генерируются стандартные строки *InChI* и *InChIKey*, это приведет к доступности содержимого баз данных для взаимодействия с другими химическими ресурсами в сети Интернет.

ИСПОЛЬЗОВАНИЕ InChIKey ДЛЯ ПОИСКА ХИМИЧЕСКИХ СТРУКТУР В БАЗЕ СТРУКТУРНЫХ ДАННЫХ ПО ХИМИИ ВИНИТИ РАН

База структурных данных по химии ВИНИТИ РАН (База СД) содержит информацию о более чем 7 млн химических структур, 4 млн химических реакций и 15 млн свойств химических соединений. Основой формирования Базы СД является система *CBASE32* [33].

В *CBASE32* сравнение химических структур осуществляется с использованием оригинального 12-байтового хэш-кода [34]. Хэш-код генерируется непосредственно из матрицы связности, минуя генерацию линейной нотации химической структуры. Но при этом химическая структура обрабатывается таким же образом, как и при генерации *InChIKey* – сначала осуществляется нормализация, затем нумерация атомов приводится к канонической. Динамический диапазон хэш-кода огромный – $8*10^{28}$. Однако из-за того, что значения хэш-кода распределены не равномерно, как для *InChIKey*, а группируются вокруг отдельных значений, в базах данных размером порядка 10^7 записей совпадения хэш-кодов для разных структур наблюдались неоднократно.

В *CBASE32* сравнение соединений по стереохимии осуществляется при решении задачи изоморфизма двух графов, при этом для идентичности вершин требуется идентичность стереохимической информации. Эта процедура требует много компьютерного времени. Поэтому, с точки зрения идентификации стереоизомеров, использование *InChI/InChIKey* технологии вместо существующей позволит заметно ускорить обработку стереохимической информации вплоть до автоматической регистрации соединений с новой стереоконфигурацией.

В ряде случаев технология обработки структур в *CBASE32* имеет преимущества. Например, в отличие от *CBASE32*, программное обеспечение *InChI Trust* не может интерпретировать проекции Фишера циклических и линейных углеводов для расчета стереоконфигурации.

В последние годы в ВИНИТИ РАН были разработаны программы, реализующие два способа предоставления информации пользователям: поиск в интерактивном режиме и поиск в автономном режиме.

Система поиска в интерактивном режиме [35] использует алгоритм бисекций по ключу *InChIKey*. Эффективность поиска в этой системе по точной структуре подтверждена на многочисленных примерах.

Другая поисковая система[2] позволяет находить информацию о химических структурах в режиме *off-line* в локальной базе данных, сформированной из Базы СД. В эту систему включена процедура генерации ключа *InChIKey*, по которому для соответствующей структуры можно вести широкий поиск информации в сети Интернет.

ЗАКЛЮЧЕНИЕ

Использование *InChIKey* для сравнения химических структур является эффективной процедурой, которую необходимо реализовывать в технологиях обработки структурной информации, связанной с формированием баз данных по химии. В частности, перевод программного обеспечения *CBASE32* на технологию *InChI* позволит решить проблему коллизий для больших баз данных, обеспечить в большинстве случаев учет стереоинформации и сделает возможным интеграцию химических баз ВИНИТИ РАН в мировую сеть.

Можно назвать следующие предпосылки для использования технологии *InChI*.

1. Поддержка *IUPAC*. Это главное преимущество *InChI*. В настоящий момент разработан *RInChI* [36, 37] – линейная нотация для описания химических реакций. В продукте предусмотрено хэширование линейной нотации в строку фиксированной длины – аналог *InChIKey*. Кроме того, различные группы *IUPAC* занимаются планированием улучшения или расширения *InChI* для описания таутомеров, металлоорганических соединений, смесей и др.

2. Свободно распространяемые алгоритмы с исходными кодами.

3. Продуманный и глубокий механизм нормализации химических структур.

4. Доступ к другим базам данных и ресурсам Интернет. Достаточно набрать *InChIKey* в поисковой строке *Google* чтобы получить детальную информацию о данной химической структуре – публикации, базы данных и др.

Главный недостаток *InChI* – игнорирование кетоенольной таутомерии при нормализации химических структур. Этот недостаток решается предобработкой структур перед генерацией линейной нотации и однозначно фиксируется.

СПИСОК ЛИТЕРАТУРЫ

1. Харари Ф. Теория графов. – М.: Мир, 1973. – 300 с.
2. Nomenclature of Organic Chemistry / eds. J. Rigaudy, S.P Klesney. – Oxford: IUPAC/Pergamon Press, 1979.
3. International Union of Pure and Applied Chemistry. Nomenclature of Inorganic Chemistry (IUPAC Recommendations 2005). – Cambridge: RSC–IUPAC, 2005.
4. Naming and Indexing of Chemical Substances for Chemical Abstracts(TM). – URL: https://www.cas.org/File%20Library/Training/STN/User%20Docs/indexguideapp.pdf.
5. Brecher J. Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature // J. Chem. Inf. Comput. Sci. – 1999. – Vol. 39(6). – P. 943–950. DOI: 10.1021/ci990062c.
6. Wiswesser W.J. How the WLN began in 1949 and how it might be in 1999 // J. Chem. Inf. Comput. Sci. – 1982. – Vol. 22 (2). – P. 88 – 93. DOI: 10.1021/ci00034a005.
7. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // J. Chem. Inf. Comput. Sci. – 1988 . – Vol. 28(1). – P. 31–36. DOI: 10.1021/ci00057a005.
8. Ash S., Cline M.A., Homer R.W., Hurst T., Smith G.B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation // J. Chem. Inf. Comput. Sci. – 1997. – Vol. 37(1). – P. 71–79. DOI: 10.1021/ci960109j.
9. Gakh A.A., Burnett M.N. Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules // J. Chem. Inf.

[2] Трепалин С.В. Свидетельство №2017613588 о государственной регистрации программы ChemDB от 22 марта 2017 г.

Comput. Sci. – 2001. – Vol. 41(6). – P. 1494–1499. DOI: 10.1021/ci000108.
10. Chi-Hsiung Lin. SEFLIN-Separate Feature Linear Notation System for Chemical Compounds // J. Chem. Inf. Comput. Sci. – 1978. –Vol. 18(1). – P. 41–47. DOI: 10.1021/ci60013a010.
11. Stein S.E., Heller S.R., Tchekhovskoi D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier // Proceedings of the 2003 International Chemical Information Conference, Infonortic. – Nimes, 2003. – P. 131–143.
12. Pletnev I., Erin A., McNaught A., Blinov K., Tchekhovskoi D., Heller S. InChIKey collision resistance: an experimental testing // Journal of Cheminformatics. – 2012. – Vol. 4. – P. 39. DOI:10.1186/1758-2946-4-39.
13. Официальный сайт группы InChI Trust. – URL: http://www.inchi-trust.org/downloads/
14. IUPAC/InChI-Trust Licence for the International Chemical Identifier (InChI) Software version 1.04, September 2011 ("IUPAC/InChI-Trust InChI Licence No.1.0"). – URL: http://www.inchi-trust.org/wp/wp-content/uploads/2014/06/LICENCE.pdf.
15. Описание формата Daylight. – URL: http://www.daylight.com/meetings/summerschool98/course/dave/smiles-intro.html.
16. Описание формата Open Eye. – URL: https://www.eyesopen.com/.
17. Heller S.R., McNaught A., Pletnev I., Stein S., Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier // Journal of Cheminformatics. – 2015. – Vol 7. – P. 22–23. DOI 10.1186/s13321-015-0068-4.
18. Lipkowitz K.B., Boyd D.B., Helson H.E. Structure Diagram Generation // Rev. Comput. Chem. – 1999. – Vol 13. – P. 313–398. DOI: 10.1002/9780470125908.ch6.
19. Trepalin S.V., Yarkov A.V., Pletnev I.V., Gakh A.A. A Java Chemical Structure Editor Supporting the Modular Chemical Descriptor Language (MCDL) // Molecules. – 2006. – Vol. 11(4). – P. 219–231. DOI:10.3390/11040219.
20. Trepalin S.V., Yarkov A.V. CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers // J. Chem. Inf. Comput. Sci. – 2001. – Vol. 41(1). – P. 100–107. DOI: 10.1021/ci000039n.
21. Документация к программному обеспечению InChI. – URL: http://www.inchi-trust.org/download/105/INCHI-1-DOC.zip.
22. Группа Marc Nicklaus. – URL: http://www.inchi-trust.org/.
23. Technical FAQ. – URL: http://www.inchi-trust.org/technical-faq/#6.4.
24. Описание алгоритма канонизации помеченного графа. – URL: http://users.cecs.anu.edu.au/~bdm/papers/pgi.pdf.
25. Weininger D., Weininger A., Weininger J.L. SMILES 2. Algorithm for Generation of Unique SMILES Notation // J. Chem. Inf. Comput. Sci. – 1989. – Vol. 29. – P. 97–101.
26. Описание алгоритма хеширования SHA-2. – URL: http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf.
27. Кнут Д. Искусство программирования. Т. 3. Сортировка и поиск / пер. с англ. – 2-е изд. – М.: Вильямс, 2007. – 824 с.
28. Список алканов с прямой цепью. – URL: https://en.wikipedia.org/wiki/List_of_straight-chain_alkanes.
29. Brink D. A (probably) exact solution to the Birthday Problem // Ramanujan Journal. –2012. – Vol 28. – P. 223–238. DOI:10.1007/s11139-011-9343-9.
30. Two isomers of spongistatin: One InChIKey. – URL: http://www-jmg.ch.cam.ac. uk/data/inchi/.
31. An InChIkey Collision is Discovered and NOT Based on Stereochemistry. – URL: http://www.chemconnector.com/2011/09/01/.
32. Reaction termochemistry data. – URL: http://webbook.nist.gov/cgi/cbook.cgi?ID=C34136000&Units=SI&Mask=8#Thermo-React.
33. Воронежева Н.И., Трепалин С.В., Чуракова Н.И., Нечаева К.С., Королева Л.М. Система представления и ввода информации о многостадийных химических реакциях с помощью программного комплекса CBASE32 // Научно-техническая информация. Сер. 2. – 2005. – №7. – С. 7–11.
34. Trepalin S.V., Yarkov A.V., Dolmatova L.M., Zefirov N.S., Finch S.A.E. WinDat: An NMR Database Compilation Tool, User Interface, and Spectrum Libraries for Personal Computers // J. Chem. Inf. Comput. Sci. – 1995. – Vol. 35(3). – P. 405–411. DOI: 10.1021/ci00025a008.
35. Нефедов О.М., Трепалин С.В., Королева Л.М., Бессонов Ю.Е. Быстрый поиск точных химических структур в больших базах данных с использованием *InChI Key* кодировки структур // Научно-техническая информация. Сер. 2. – 2013. – № 12. – С. 27–33.
36. Grethe G., Goodman J.M., Allen C. International chemical identifier for reactions (RInChI) // J. Cheminform. – 2013. – Vol. 5. – P. 45. DOI: 10.1186/1758-2946-5-45.
37. Документация по RInChI. – URL: http://www.inchi-trust.org/download/RInChI/RInChI-V1-00.zip.

Материал поступил в редакцию 12.03.18.

Сведения об авторах

ТРЕПАЛИН Сергей Владимирович – кандидат химических наук, ведущий научный сотрудник ФГБУН Институт физиологически активных веществ РАН
e-mail: trep@chemical-block.com

БЕССОНОВ Юрий Ефимович – кандидат технических наук, ведущий научный сотрудник ВИНИТИ РАН
e-mail: bessonov-ye@rambler.ru

ФЕЛЬДМАН Борис Семенович – старший научный сотрудник ВИНИТИ РАН
e-mail: bsf@inbox.ru

ЧУРАКОВА Наталия Исааковна – кандидат химических наук, зав. Отделом исследований и обработки структурной химической информации ВИНИТИ РАН
e-mail: nichurak@rambler.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 811:161.1’322.3

И.В. Аблов, В.Н. Козичев, А.В. Ширманов, Ал-др А. Хорошилов, А.А. Хорошилов

Средства машинной грамматики русского языка (по Г.Г. Белоногову)

Рассматриваются принципы и методы создания программных и декларативных средств машинной грамматики русского языка, которые базируются на оригинальных алгоритмах, разработанных научным коллективом сотрудников ВИНИТИ, 27 ЦНИИ МО РФ и Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН). Декларативные средства, представляющие собой комплекс словарей и грамматических таблиц в машинной форме, создавались на основе широкомасштабных исследований больших объемов политематической текстовой информации (измеряемой десятками миллионов слов) лингво-статистическими методами. В состав комплекса декларативных средств входят грамматические таблицы и машинные словари, включающие основные типы словоизменительных и словообразовательных трансформаций, а также представительные словари основ слов. На базе использования этих декларативных средств разработаны уникальные алгоритмы машинной грамматики русского языка. Описанные средства в настоящее время широко используются для решения сложных задач автоматической обработки и анализа смыслового содержания текстовой информации в ряде промышленных информационных систем.

Ключевые слова: *русский язык, машинная грамматика, морфологический анализ, словоизменение, словообразование, программные средства, декларативные средства, автоматическая нормализация слов*

ВВЕДЕНИЕ

Начало исследований и разработок по созданию описанной ниже машинной грамматики русского языка нужно отнести к концу 1950-х – началу 1960-х гг., когда в Центральном научно-исследовательском институте № 27 Министерства обороны СССР доктором технических наук, полковником Г.Г. Белоноговым впервые был составлен частотный словарь русских слов по текстам военного назначения. Этот словарь послужил основой для разработки концепции машинной грамматики и создания на его основе системы флективных классов слов русского языка, которые позволили систематизировать основные типы словоизменительных парадигм грамматических классов слов. В дальнейшем в 1980-х годах во Всероссийском институте научной и технической информации (ВИНИТИ) эта концепция была апробирована на современных научно-технических текстах баз данных ВИНИТИ. В результате выполненных в ВИНИТИ исследований был составлен машинный пятисотмиллионный словарь основ слов, на основе которого были разработаны компактные грамматические таблицы для машинных процедур автоматического анализа текстов.

Необходимо отметить, что аналогичными исследованиями в области машинной грамматики русского языка в 1960-х годах прошлого столетия занимались такие известные ученые-лингвисты как И.А. Мельчук, Ю.Д. Апресян, А.А. Зализняк и др. В рамках разработанной И.А. Мельчуком и его последователями фундаментальной модели «Смысл ⇔ Текст» естественного языка был разработан морфологический компонент модели [1], который по замыслу авторов представлял собой попытку единообразного определения традиционных морфологических понятий и исчисления грамматических категорий в ряде языков мира. Практическим результатом этого научного направления явилось создание в 1977 г. А.А. Зализняком грамматического словаря русского языка, включающего около 100 тыс. слов современного русского языка с полным морфологическим описанием [2]. Этот основополагающий труд по морфологии русского языка базировался на системном подходе к описанию

грамматических парадигм, включающих не только изменение буквенного состава слов, но и ударения в словах. Словарь отражал (с помощью специальной системы условных обозначений) современное словоизменение (склонение существительных, прилагательных, местоимений, числительных и спряжение глаголов) и является обратным словарём, где слова упорядочены по последним буквам.

Такие машинные словари предназначены для использования в процедурах морфологического анализа для автоматического определения морфологических характеристик слов: установления границы между основой и окончанием, а также определения набора их характеристик, соответствующих грамматической форме слова. Эти средства позволяют устанавливать смысловое тождество членов словоизменительных и словообразовательных парадигм. Средства машинной грамматики базируются на машинных словарях и на оригинальных алгоритмах, разработанных научным коллективом сотрудников ВИНИТИ, 27 ЦНИИ МО РФ и Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН). Разработанный программный комплекс машинной грамматики русского языка в настоящее время успешно используется в составе различных промышленных информационных систем в ряде федеральных органов власти, учебных и научно-исследовательский институтов и крупных информационных центров Российской Федерации.

СИСТЕМА ФЛЕКТИВНЫХ КЛАССОВ РУССКИХ СЛОВ

Общеизвестно, что русский язык относится к типу флективных языков, в которых основным средством словоизменения являются окончания, а средством словообразования – приставки и суффиксы. Ввиду того, что словообразование с помощью приставок существенно изменяет смысл слов, а словообразование с помощью суффиксов –незначительно модернизируют их смысл, мы рассматриваем только суффиксальное словообразование.

Слова русского языка традиционно разделяются на ряд лексико-грамматических классов. Основные классы слов и их мнемоническое обозначение приведены в табл. 1.

Обычно в составе слова различают несколько типов морфем: корневые морфемы (корни), префиксы (приставки) и суффиксы (морфемы, стоящие после корня). Основную смысловую нагрузку несет корень, а префиксы и суффиксы выступают в роли модификаторов смысла. Например, в слове *выступающий* можно выделить пять морфм: *вы-ступ-а-ющ-ий*. Здесь морфема *ступ* – корень слова, морфема *вы* – префикс, морфемы *а, ющ* и *ий* – суффиксы (суффикс *ий* является грамматическим окончанием).

Таблица 1

Лексико-грамматические классы слов и их мнемонические обозначения

№	*Грамматический класс слова*	*Мнемоническое обозначение*
1	Существительное	N
2	Прилагательное	A
3	Глагол	V
4	Субстантивированное прилагательное	S
5	Местоимение	M
6	Числительное	C
7	Союз	&
8	Междометие	!
9	Наречие	Y
10	Предлог	F
11	Причастие	W
12	Деепричастие	D

В процессе функционирования языка слова могут приобретать различные формы в различных контекстных окружениях. Это могут быть формы словоизменения и словообразования. Граница между ними условная, и различные авторы устанавливают ее поразному. Можно, например, считать, что формы склонения существительных и прилагательных, формы спряжения глагола настоящего и будущего времени, формы изменения глаголов прошедшего времени, кратких прилагательных и кратких причастий по родам и числам являются формами словоизменения, а все остальные трансформации слов – формами словообразования. Мы придерживаемся именно такой позиции. Фрагмент списка окончаний слов приведен в табл. 2, их число составляет 77 различных окончаний слов [3].

Изменения форм слов могут носить различный характер. Они могут быть связаны как с изменением основы слова, так и с изменением его окончания. Изменение буквенного состава основ имеет место, например, в следующих парах форм слов: *сижу – сидишь*, *шел – шли*, *тренировка – тренировок*, *нес – несли*, *кто – кого*, *время – времени*, *судно – суда*, *человек – люди*. Изменение грамматических окончаний является основным способом образования различных форм слов при изменении их рода, числа, падежа и лица. В русском языке этот способ словоизменения используется как самостоятельно, так и в сочетании с изменением основ слов.

Таблица 2

Фрагмент списка грамматических окончаний русских слов

Грамматические окончания русских слов
+ а ам ами ась ат атся ... им ими имися имся ись ит ите итесь ится их ихся ишь ишься … ого ое ой ом … юсь ют ются ою я ям ями ят ятся ях яя
Знак “+” – обозначает отсутствие буквенного кода окончания

Суффиксы и префиксы бывают не во всех словах. Например, в слове *переход* есть префикс *пере*, но отсутствуют суффиксы; в слове *ходящий* есть сочетание суффиксов *ящий*, но нет префиксов; а в слове *ход* нет ни суффиксов, ни префиксов, есть только одна корневая морфема – *ход*. В русском языке могут встречаться слова и без формально обозначенной корневой морфемы – слова с так называемой "нулевой" корневой морфемой. Примером тому может служить слово *вынуть*. Здесь буквосочетание *вы* является префиксом, а буквосочетание *нуть* – сочетанием суффиксов (подобно словам *прыг-нуть* и *шаг-нуть*).

В основу рассматриваемой машинной грамматики положена система флективных классов русских слов, в которой в форме компактной таблицы представлены основные типы словоизменительных парадигм лексико-грамматических классов русских слов. Эта таблица была построена на базе многолетних широкомасштабных исследований русских научно-технических текстов, выполненных коллективом исследователей под руководством Г.Г. Белоногова в 1960-х гг. [3]. Было установлено, что по характеру изменения грамматических окончаний (флексий) и по своей синтаксической функции русские слова могут быть разбиты на ряд классов, которые получили название флективных. Флективные классы изменяемых слов выделяются на основе анализа их синтаксической функции и систем падежных, личных и родовых окончаний. Классы неизменяемых слов – только по синтаксической функции. Фрагменты таблицы флективных классов слов приведены в табл. 3.

По своей функции изменяемые слова объединены в следующие группы: 1) существительные; 2) прилагательные; 3) глаголы в личной форме; 4) глаголы прошедшего времени, краткие прилагательные и краткие причастия: 5) количественные числительные. Группа "существительные", в свою очередь, разбивается на несколько подгрупп, выделенных по признакам рода и одушевленности (для существительных мужского и женского рода). В каждой группе и подгруппе слова распределены по флективным классам.

Определение принадлежности изменяемого слова к группе или подгруппе обычно не вызывает затруднений, так как в основу принятого здесь разделения на группы и подгруппы положена традиционная классификация слов. Следует лишь учитывать, что, наряду с полными прилагательными, к группе "прилагательные" отнесены также полные причастия, порядковые числительные, субстантивированные прилагательные, а также количественное числительное один. При выделении окончания слова возвратные частицы *ся*, *сь* и внутренний мягкий знак (мягкий знак, стоящий между основой и ненулевым окончанием слова) опускаются.

Некоторые классы существительных мужского и женского рода имеют одинаковые окончания во всех формах, принятых в качестве типичных, хотя другие их формы не совпадают. Иллюстрацией могут служить пары слов: *огонь – путь, перебой – санаторий, сосед – врач, нутрия – швея, грань – речь, линия – галерея.*

Наряду с рассмотренными выше способами варьирования форм слов, которые мы назвали способами словоизменения, в практике речевого общения широко используются и способы словообразования. Словообразовательные трансформации слов связаны, прежде всего, с изменением состава их префиксов и суффиксов. При этом может иметь место также чередование гласных и согласных в корневых морфемах (например, у пар слов *проводить – проведение, относиться – отношение, убедившийся – убежденный, проношу – пронесли*). Перечень наиболее часто встречающихся суффиксов и сочетаний суффиксов приведен в табл. 4, их число составляет 1591 суффикс или сочетание суффиксов. В этой таблице каждому из них поставлен в соответствие номер флективного класса. Это связано с тем, что суффиксы и их сочетания, имеющие одинаковый буквенный состав, но относящиеся к различным флективным классам (совместимые с различными наборами окончаний), считаются разными. Например, суффиксы "*н*" в словах *главный* и *отрывной* и суффиксы "*ов*" в словах *портовый* и *годовой* – разные суффиксы.

Таблица 3

Фрагменты таблицы флективных классов русских слов

Существительные (N)	
№ ФК	Слово-представитель класса и набор грамматических окончаний
001	Телефон – *+ а у + ом е ы ов ам ы ами ах*
002	Тираж *+ а у + ом е и ей ам и ами ах*
	. . . . .
	Прилагательные (A)
103	Главный – *ый ого ому ый ым ом ая ой ой ую ой ой ое ого ому ое ым ом ые ых ым ые ыми ых*
104	Передний – *ий его ему ий им ем яя ей ей юю ей ей ее его ему ее им ем ие их им ие ими их*
	. . . . .
	Глаголы личной формы (V)
116	Делать – *ю ешь ет ем ете ют*юсь ешься ется емся етесь ются*
117	Строить – *ю ишь ит им ите ят*юсь ишься ится имся итесь ятся*
	. . . . .

Таблица 4

Фрагмент таблицы суффиксов (сочетаний суффиксов) русских слов

Суффиксы (сочетания суффиксов) с номерами флективных классов
а/004 а/116 а/152 абельн/103 абельн/126 абельно/152 абельност/055 … ец/001 ец/011 ец/021 ец/032 ец/057 ечек/060 ечк/060 ечн/103 … иру/116 ируемост/055 ируй/143 ируйся/143 ируйте/143 ируйтесь/143 ирующ/105 ируя/152 …

МОРФОЛОГИЧЕСКИЙ СЛОВОИЗМЕНИТЕЛЬНЫЙ АНАЛИЗ РУССКИХ СЛОВ

Морфологический словоизменительный анализ – это лингвистическая процедура, предназначенная для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения ряда процедур автоматической обработки текстовой информации, таких например, как процедур морфологического синтеза слов, синтаксического анализа текстов, синтаксического синтеза текстов и концептуального анализа. На основе этой процедуры обеспечивается возможность установления смыслового тождества различных форм слов словоизменительной парадигмы. Морфологический словоизменительный синтез предназначен для генерации заранее заданных форм слов словоизменительной парадигмы.

Традиционно процедуры морфологического анализа включали два этапа анализа слов – точный анализ, базирующийся на использовании словарей основ слов и приближенный, предназначенный для анализа слов, отсутствующих в словарях [2,4]. Точный анализ при всех его преимуществах по обеспечению возможности назначения правильных грамматических характеристик слов требовал большого объема памяти и значительного числа итераций при поиске основы анализируемого слова в словаре основ и операции проверки на совместимость найденной основы со списком окончаний. При этом в случае отсутствия в словаре основы анализируемого слова приходилось дополнительно подключать процедуру приближенного анализа. Такие процедуры морфологического анализа не отличались большим быстродействием. Авторы настоящей статьи под руководством Г.Г. Белоногова разработали быстродействующий алгоритм приближенного анализа русских слов, который по точности назначения грамматических признаков не уступал точному анализу, выполняемому по реальному словарю объемом более миллиона слов [3].

Основная идея разработанного авторами алгоритма базировалась на гипотезе, что *в русском языке объективно существует сильная корреляция между конечным буквосочетанием слов и их грамматическими характеристиками.* Реализация этой гипотезы позволила свести огромные словари словоформ в две небольшие грамматические таблицы и применить быстродействующий алгоритм с минимальным числом итераций. Основными грамматическими таблицами являлись таблица конечных буквосочетаний слов (таблица КБС), предназначенной для обработки слов по методу аналогии и таблица служебных и коротких слов (СКС), включающая в свой состав “служебные” и короткие слова, словоизменения которых имели аномальные трансформации.

При разработке этих таблиц были проведены исследования больших объемов текстовой информации лингво-статистическими методами с целью выявления всех возможных словоизменительных и словообразовательных трансформаций слов. В результате этих исследований были разработаны эффективные методы сжатия больших словарных ресурсов в компактные грамматические таблицы для процедур анализа и синтеза русских слов. В основу этих методов были положены принципы лингвистической аналогии. Приводятся основные грамматические словари и таблицы для словоизменительного анализа слов:

Таблица КБС – таблица конечных буквосочетаний слов, предназначенная для установления по конечному буквосочетанию слова его грамматических характеристик: длины окончания, флективного класса и модели управления. Эта таблица создана на основе лингвистического анализа словоизменительных трансформаций более полумиллиона словоформ. В ней представлены основные типы конечных буквосочетаний, позволяющие с высокой вероятностью назначать анализируемым словам основные грамматические характеристики. В машинном представлении этой таблицы прямой порядок следования буквенного состава конечных буквосочетаний преобразован в их обратный порядок следования. Таблица включает 27547 элементов. В табл. 5 приведены фрагменты таблицы КБС в двух представлениях – в прямом и инверсном порядке.

Словарь СКС – словарь “служебных” и коротких слов включает “служебные” (предлоги, союзы, местоимения и др.), короткие слова (менее 5-ти букв) и супплетивные формы слов, а также слова, словоизменительные трансформации которых имели аномальные реализации. Объем словаря – 78356 словарных статей. В табл. 6 приведены фрагменты словаря СКС. В словаре для каждого исходного слова с назначенной грамматической информацией приведены все формы слов его словоизменительной парадигмы. Строгий позиционный порядок их следования позволяет соотносить эти формы с набором их грамматической информации.

Таблица ФКГИ – таблица, устанавливающая однозначное соответствие между грамматическими характеристиками слов: буквенным кодом окончаний слов, их флективным классом и грамматическими признаками: родом, числом, падежом и лицом этого слова. Таблица создана путем лингвистического ана-

лиза полумиллиона русских слов и представляет собой компактную таблицу объемом 1584 элемента (см. табл. 7). В каждой колонке таблицы в строку расположены: буквенный код окончания слова (на первой слева позиции), трехзначный цифровой индекс его флективного класса (на второй слева позиции) и четырехзначный цифровой индекс грамматических признаков слова: род, число, падеж, лицо. Каждая цифра четырехзначного представления грамматических признаков имеет следующее значение:

Род: 0 – не определен, 1 – мужской, 2 – женский , 3 – средний;

Число: 0 – не определено, 1 – единственное, 2 – множественное;

Падеж: 0 – не определен, 1 – именительный, 2 – родительный, 3 – дательный,
4 – винительный, 5 – творительный, 6 – предложный;

Лицо: 0 – не определено, 1 – 1-ое лицо, 2 – 2-ое лицо , 3 – 3-ье лицо.

Таблица 5

Фрагменты таблицы конечных буквосочетаний слов

Прямой порядок конечных буквосочетаний	Инверсный порядок конечных буквосочетаний
……	……
……*абр 01/001/01*	*рба 01/001/01*
……*абу 00/145/01*	*уба 00/145/01*
……*абуку 01/056/01*	*укуба 01/056/01*
……*абур 01/001/01*	*руба 01/001/01*
……*абут 01/056/01*	*туба 01/056/01*
……*абы 01/044/01*	*ыба 01/044/01*
……*абь 01/056/01*	*ьба 01/056/01*
……	……
……*авес 01/001/01*	*сева 01/001/01*
……*авет 01/126/01*	*сева 01/001/01*
……*авеу 01/042/01*	*уева 01/042/01*
……*авеци 01/126/01*	*ищева 01/126/01*
……*авецм 01/042/01*	*мцева 01/042/01*
……*авецне 01/126/01*	*енцева 01/126/01*
……	……
……*авецнероле 01/042/01*	*елоренцева 01/042/01*
……*авечу 01/126/01*	*учева 01/126/01*
……*авечур 01/042/01*	*ручева 01/042/01*
……*авешыбй 01/001/01*	*йбышева 01/001/01*
……*авещ 01/042/01*	*щева 01/042/01*
……*авеьв 01/053/01*	*вьева 01/053/01*
……*авеьд 01/042/01*	*дьева 01/042/01*
……	……
……*алсы 01/001/01*	*ысла 01/001/01*
……*алу 01/056/01*	*ула 01/056/01*
……*алуа 01/001/01*	*аула 01/001/01*
……*алуб 01/056/01*	*була 01/056/01*
……*алубй 01/001/01*	*йбула 01/001/01*
……*алуд 01/056/01*	*дула 01/056/01*
……	……

Таблица 6

Фрагменты словаря “служебных” и коротких слов

Исходная форма	Символ слова	Формы слов словоизменительной парадигмы
люди/01/307/01/1210/		*человек человека человеку человека человеком человеке люди людей людям людей людями людях*
меня/00/307/02/1120,1140	*N*	*я, меня, мне, меня, мной, мне, мы, нас, нам, нас, нами, нас*
мне/00/307/02/1130,1160	*N*	*я, меня, мне, меня, мной, мне, мы, нас, нам, нас, нами, нас*
мы/00/307/02/1210	*N*	*я, меня, мне, меня, мной, мне, мы, нас, нам, нас, нами, нас*
нам/00/307/02/1230	*N*	*я, меня, мне, меня, мной, мне, мы, нас, нам, нас, нами, нас*
нами00/307/02/1250	*N*	*я, меня, мне, меня, мной, мне, мы, нас, нам, нас, нами, нас*
человек /00/307/02/1210	*N*	*человек человека человеку человека человеком человеке люди людей людям людей людями людях*

Таблица 7

Фрагменты таблицы флективных классов и грамматических признаков

Соответствие буквенного кода окончаний, флективного класса и грамматических признаков слов (род, число падеж, лицо)		
. . . . .	. . . . .	. . . . .
+ 001 1110	ей 114 2120	у 002 1130
+ 001 1140	ей 114 2130	у 006 1130
+ 002 1110	ей 114 2150	у 007 1130
+ 002 1140	ей 114 2160	у 010 1130
+ 006 1110	ей 115 2120	у 010 1160
+ 006 1140	ей 115 2130	у 011 1130
+ 007 1110	ей 115 2150	у 014 1130
+ 007 1140	ем 020 1150	у 032 1130
+ 015 1140	ем 025 1150	у 033 1140
+ 015 1220	ем 032 1150	у 037 1130
. . . . .	. . . . .	. . . . .
Знак “+” – обозначает отсутствие окончания		

Таблица 8

Результаты работы морфологического анализа русских слов

Буквенный код исходного слова	Символ класса	Грамматич. характеристики слова: длина оконч., флект. класс, мод. упр, род, число, падеж, лицо
российская	*A*	*02/106/01/2110*
сторона	*N*	*01/056/01/2110*
поставила	*L*	*01/125/34/2100*
вопрос	*N*	*00/001/01/1110,1140*
о	*F*	*00/164/46/0400,0600*
неполном	*A*	*02/103/01/1160,3160*
исполнении	*N*	*01/073/02/3130*
резолюции	*N*	*01/061/01/2120,2130,2160,2210,2240*

По любым двум грамматическим характеристикам словарной статьи этой таблицы можно однозначно установить третью характеристику. Например, имея информацию о флективном классе слова и окончании, можно однозначно установить набор грамматических признаков: род, число, падеж, лицо.

Порядок обработки слов алгоритмом морфологического словоизменительного анализа выполняется следующим образом: вначале производится поиск анализируемого слова в таблице “служебных” и коротких слов и, если оно там находится, ему назначается соответствующий набор грамматических признаков найденного слова. Если это слово не было обнаружено в словаре СКС, то производится инверсия его буквенного состава и выполняется поиск на наибольшее совпадение конечного буквосочетания слова с буквосочетанием одного из элементов таблицы КБС. После установления такого буквосочетания анализируемому слову назначаются его грамматические характеристики. Недостающая грамматическая информации устанавливается по таблице ФКГИ.

Приведем алгоритм морфологического анализа слов русского языка.

Алгоритм 1. (алгоритм морфологического словоизменительного анализа русских слов).

Шаг 1. Выполняется поиск анализируемого слова на полное его совпадение в словаре СКС. В случае успешного поиска слову назначается по этому словарю грамматическая информация и выполняется переход к шагу 5. В случае отсутствия этого слова в словаре – переход к шагу 2.

Шаг 2. Производится инверсия буквенного состава анализируемого слова и выполняется поиск конечного буквосочетания анализируемого слова на наибольшее совпадение с одним из элементов словаря КБС. Переход к шагу 3.

Шаг 3. Исходной форме анализируемого слова назначаются грамматические характеристики совпавшего элемента словаря КБС: мнемонический символ класса слова, флективный класс, длина окончания и модель управления. Переход к шагу 4.

Шаг 4. На основании информации о флективном классе и буквенном коде окончания слову по словарю ФКГИ назначается набор грамматических признаков: (род, число, падеж, лицо). Переход к шагу 5.

Шаг 5. Преобразование полученных результатов в структуру метаданных.

Результаты работы морфологического анализа русских слов приводятся в табл. 8.

В этой табл. 8 используются мнемонические обозначения классов слов из табл.1, а значения цифр четырехзначных представлений грамматических признаков – табл. 7.

МОРФОЛОГИЧЕСКИЙ СЛОВОИЗМЕНИТЕЛЬНЫЙ СИНТЕЗ РУССКИХ СЛОВ

Морфологический словоизменительный синтез русских слов – это лингвистическая процедура, обеспечивающая возможность автоматической генерации требуемой формы слова словоизменительной парадигмы на основе исходной формы слова и заданного набора грамматической информации. Здесь, казалось бы, напрашивается достаточно простое решение – к текстовой словоизменительной основе нужно присоединить грамматическое окончание, соответствующее заданной грамматической информации. Но, к сожалению, решение этой задачи осложняется тем, что при решении задачи синтеза нужно преодолеть проблемы чередования гласных в корневых морфемах и наличия у некоторых слов их супплетивных форм.

Первая проблема была решена путем выявления и анализа основных типов трансформаций в корневых морфемах слов при их словоизменении. В процессе этого анализа были установлены грамматические классы слов, у которых такие трансформации возможны, а также были установлены конечные буквосочетания основ, при которых они встречаются. Этот анализ также показал, что у таких слов (с чередованием в корневой основе) существует два типа основ – канонические и вариантные. Под канонической основой понимается та основа, которая характерна для канонических форм слов. Под вариантной формой – та основа, которая характерна для форм слов, отличных от канонической. В таблице соответствия грамматических признаков слов типу основы (ГПТО) указывается: какие формы слов словоизменительной парадигмы содержат каноническую форму основы, а какие формы – вариантную. В ней для каждого набора грамматической информации (значения цифр наборов информации этой таблице идентичны обозначениям табл. 6, буквами *К* и *В* указан тип основы (*К* – каноническая, *В* – вариантная). Фрагмент таблицы соответствия грамматических признаков слов типу основы приведен в табл. 9.

С помощью таблицы списка подстановок (СП) можно определить по флективному классу слову и конечному буквосочетанию основы наличие или отсутствие чередования основы, а также механизм реализации трансформации основы при синтезе заданной формы слова. В этой таблице для каждого флективного класса указывается список конечных буквосочетаний основ, которые определяют чередование и механизм его реализации: сколько символов от конца основы нужно отбросить и какие символы нужно присоединить вместо удаленных символов. Ввиду того, что установление типа основы производится по ее инверсному представлению, конечные буквосочетания и подстановки в табл. 10 также представлены в инверсном виде.

Таблица 9

Фрагмент таблицы соответствия грамматических признаков слов типу основы

№ ФК	Слово-представитель и наборы грамматических признаков слов совместимые с типами основ
001	***ветер****- 1110-К, 1120-В, 1130-В, 1140-К, 1150-В, 1160-В, 1210-В, 1220-В, 1230-В, 1240-В, 1250-В, 1260-В*
006	***уголок****- 1110-К, 1120-В, 1130-В, 1140-К, 1150-В, 1160-В, 1210-В, 1220-В, 1230-В, 1240-В, 1250-В, 1260-В*
060	***заготовка****- 1110-К, 1120-К, 1130-К, 1140-К, 1150-К, 1160-К, 1210-К, 1220-В, 1230-К, 1240-К, 1250-К, 1260-К*

Таблица 10

Фрагмент таблицы списка подстановок

№ ФК	Номер списка подстановок, тип основы, правила подстановок (идентификатор конечных буквосочетаний основы, кол. отделяемых букв, заменяемые символы
001	***352_К****# цлав-2,ец/табед-1,т/лсымс-1,л/ртсок-1,ер/ньлек-1,н/ырыс-0,е/цпищ-1,ц/церт-1,ц/цбуг-1,ц/тпиг-1,ет/ртев-1,ер/рташ-1,ер/рдак-1,р/рвок-1,ер/лпеп-1,ел/мйат-1,м/лзу-1,ел/цса-1,ц/цьл-2,ец/лгу-1,ол/лси-1,ел/тбе-1,ет/лсы-1,ел/сво-1,ес/лто-1,ел/лхе-1,ол/мйа-2,ем/ргу-1,ор/ньл-2,ен/цр-1,ец/вш-1,ов/цп-1,ец/цн-1,ец/цз-1,ец/цд-1,ец/цв-1,ец/лс-1,ел/цб-1,ец/нс-1,он/*
	544_В*# ретсок-2,р/лохеч-2,л/тебер-2,т/север-1,с/севон-1,с/церут-1,ц/севол-1,с/севок-1,с/севод-1,с/севов-1,с/леток-2,л/ретев-2,р/реташ-2,р/ревок-2,р/лесик-2,л/нелч-1,н/нелт-1,н/нелп-1,н/нело-1,н/нелк-1,н/носо-2,н/носу-2,н/рогу-2,р/сево-2,с/тепш-1,т/ворк-1,в/воро-1,в/ворт-1,в/воше-1,в/церт-1,ц/дела-1,д/дело-1,д/делп-1,д/делс-1,д/лезу-2,л/церк-1,ц/лепм-1,л/церб-1,ц/лесы-2,л/логу-2,л/цепи-2,ц/нела-1,н/неле-1,н/нели-1,н/цен-2,ц/цел-2,ьц/цеп-2,ц/цер-2,ц/цет-2,ц/цез-2,ц/цев-2,ц/цеб-2,ц/теп-2,т/нел-2,ьн/меа-2,йм/лес-2,л/леп-2,л/дел-2,ьд/вош-2,в/вор-2,в/*
…	……

Таблица 11

Результаты работы морфологического словоизменительного синтеза

Буквенный код исходного слова	Символ класса исх. слова	Грамматическая информация о слове	Грамматическая информация для синтеза слова	Буквенный код синтезируемого слова
российская	*A*	*02/106/01/2110*	*1120*	*российского*
сторона	*N*	*01/056/01/2110*	*2130*	*стороне*
поставила	*L*	*01/125/34/2100*	*1200*	*поставили*
вопрос	*N*	*00/001/01/1110,1140*	*1250*	*вопросами*
о	*F*	*00/164/46/0400,0600*	*0000*	*о*
неполном	*A*	*02/103/01/1160,3160*	*2110*	*неполная*
исполнении	*N*	*01/073/02/3130*	*3110*	*исполнение*
резолюции	*N*	*01/061/01/2120,2130,2160,2210, 2240*	*2150*	*резолюцией*

Вторая проблема – проблема замены исходных форм слов на их супплетивные формы выполняется с помощью словаря СКС. Отличительным признаком супплетивной формы является флективный класс №307. В этом словаре позиционный порядок следования каждой супплетивной форме слова и грамматическая информация исходной формы слова позволяет однозначно соотнести синтезирующую информацию с позицией требуемой супплетивной формой.

Далее приведен алгоритм морфологического словоизменительного синтеза. На вход этой процедуры подается исходное текстовое слово, результаты его морфологического анализа и набор грамматической информации, задающий синтезируемую форму слова.

Алгоритм 2. (алгоритм морфологического синтеза русских слов).

Шаг 1. По номеру флективного класса устанавливается: имеются ли в словоизменительной парадигме исходного слова супплетивные формы. Если супплетивных форм нет, то переходим к шагу 3. Если супплетивные формы имеются, то определяется: соответствует ли форма исходного слова синтезируемой форме слова. Если соответствует, то переходим к шагу 7. Если не соответствует, то переходим к шагу 2.

Шаг 2. По исходной форме и задающей грамматической информации из словаря СКС выбирается соответствующая форма. Переход к шагу 7.

Шаг 3. Устанавливается: имеется ли чередование в основе исходного текстового слова. В случае отсутствия чередования переходим к шагу 5. Если имеется чередование, то переходим к шагу 4.

Шаг 4. Если в основе устанавливается факт наличия чередования, то определяется: соответствует ли тип исходного слов типу синтезируемого слова. Если соответствует, то переходим к шагу 6. Если не соответствует, то переходим к шагу 5.

Шаг 5. Если устанавливается, что тип основы исходного слова не соответствует типу основы синтезируемого слова, выполняется изменение конечного буквосочетаний исходной основы. После трансформации основы переходим к шагу 6.

Шаг 6. Если установлено, что основа исходного слова соответствует основе синтезируемого слова, то необходимо к словоизменительной основе присоединить требуемое грамматическое окончание. Для этого в соответствии с заданным набором грамматических признаков и на основании информации о флективном классе и буквенном коде основы слова к его концу присоединяется грамматическое окончание, установленное по словарю ФКГИ. Переход к шагу 7.

Шаг 7. Преобразование полученных результатов в структуру метаданных.

Результаты работы морфологического словоизменительного синтеза приведены в табл.11.

В табл. 11 используются мнемонические обозначения классов слов из табл.1, а значения цифр четырехзначных представлений грамматических признаков из табл. 7.

МОРФОЛОГИЧЕСКИЙ СЛОВООБРАЗОВАТЕЛЬНЫЙ АНАЛИЗ РУССКИХ СЛОВ

Морфологический словообразовательный анализ слов – это лингвистическая процедура, предназначенная для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения различных процедур автоматической обработки текстовой информации, требующих установлении смыслового тождества слов на уровне словообразования.

Необходимо отметить, что словообразовательный анализ значительно сложнее словоизменительного анализа (который, как правило, является составной частью словообразовательного анализа). При этом также более сложно реализуется процедура словообразовательного синтеза. Для реализации этих процедур существенно расширен состав декларативных средств. В их состав были включены словари словообразовательных трансформаций слов русского языка и представительные словари основ слов.

Словарь СоКС – словарь словообразовательных классов слов, создан на основе анализа словообразовательных трансформаций полумиллиона словоизменительных основ слов [5]. В процессе этого анализа

были выявлены основные типы словообразовательных трансформаций, зафиксированные словаре СоКС. Текущая версия словаря СоКС включает 1380 классов. В табл. 12 приведены фрагменты словаря СоКС. Каждому словообразовательному классу поставлено в соответствие слово – представитель класса и перечень суффиксов (сочетаний суффиксов), входящих с состав класса.

На основе словаря СоКС были разработаны другие формы его представления. В частности, был разработан инвертированный словарь словообразовательных классов слов (словарь СоКС_И) [6]. В этом словаре каждому суффиксу (сочетанию суффиксов) поставлены в соответствие списки номеров словообразовательных классов, в состав которых эти суффиксы входят. Фрагменты словаря СоКС_И представлены в табл. 13. Объем словаря составляет 1380 словарных статей.

Словарь СОС – словарь словообразовательных основ слов, создан на основе статистического анализа и автоматизированной обработки больших объемов текстов по широкому спектру тематических областей [7,8]. Словарь содержит все типы словоизменительных и словообразовательных трансформаций и представляет собой словарь словоформ, в которых определены словоизменительные и словообразовательные основы. Общий объем словаря 1,2 млн словоизменительных основ, содержащих 286 тыс. словообразовательных основ русских слов. Этот словарь используется в качестве исходных данных для различных исследований и реализаций процедур анализа текстов.

Таблица 12

Фрагменты словаря словообразовательных классов слов

№ СоКС	Слово-представитель класса и перечень суффиксов (сочетаний суффиксов) класса
0002	***Разброс*** *+*001 +*124 а*116 ав*152 авш*105 авши*152 ай*143 айте*143 ал*125 ан*126* аться*144 анн*103 ать*144
0003	***Изрез-анность*** *а*116 ав*152 авш*105 авши*152 ал*125 ан*126 анн*103 ающ*105 анност*055 ать*144*
	. . . . .
0008	Собир-ательный *а*116 авш*105 аем*103 ай*143 айся*143 айте*143 айтесь*143 ал*125 ател*027 ательн*103 ани*073 ать*144 аться*144 ательно*152 ая*152 аясь*152 ающ*105*
	. . . . .
0036	***Обращ-аются*** *+*124 а*116 авш*105 аем*103 аемост*055 ай*143 айся*143 айте*143 айтесь*143 ал*125 ающ*105 ать*144 аться*144 ен*126 ая*132 аясь*152 енност*055 ени*073 енн*103*
	. . . . .
0039	***Замеч-ать*** *+*124 а*116 авш*105 аем*103 ай*143 айте*143 ал*125 ающ*105 ани*073 ать*144 енн*103 ая*152 ен*126*

Таблица 13

Фрагменты инвертированного словаря словообразовательных классов слов

Суффикс (сочетание суффиксов)	Перечень СоКС (с данным суффиксом (сочетанием суффиксов))
…….	……
*а*116*	*0002 / 0010 / 0025 / 0050 / 0054 / 0066.... 0537 / 0574 / 0576 / 0588 / 0590 / 0607 ... 0848 / 0855 / 0859 / 0860 ... 1039 / 1054 / 1090 / 1101 ... 1332 / 1334*
*ать*144*	*0002 / 0010 / 0025 / 0050 / 0054 / 0066.... 0537 / 0574 / 0576 / 0588 / 0590 / 0607 ... 0848 / 0855 / 0859 / 0860 ... 1039 / 1054 / 1090 / 1101 ... 1332 / 1334*
…….	……
*ец*001*	*1178 / 1196*
*ец*011*	*1195*
*ец*021*	*0433 / 0437 / 1183 / 1192 / 1201*
*ец*032*	*1174 / 1175 / 1181 / 1182 / 1184 / 1188 / 1190 / 1191 / 1204 / 1213*
*ец*057*	*1189*
*ец*001*	*1178 / 1196*
…….	……

Таблица 14

Фрагмент словаря словарь словообразовательных основ слов

Буквенный код слов	Грамматическая информация
……..	……..
автомат-изаци-я	05/02/061/2
автомат-изационн-ый	08/02/103/01
автомат- изировавш-ый	09/02/105/24
автомат-изировал-+	08/00/125/24
автомат-изирован-+	08/01/126/01
автомат-изировани-е	09/01/073/24
автомат-изировать-+	09/01/144/24
автомат-изироваться-+	11/01/144/01
автомат-изиру- ет	05/02/116/24
автомат-изируем-ый	07/02/103/5
автомат-изирующ -ий	08/02/105/24
автомат-изируя	06/ 01/152/24
автомат-изм-+	03/ 01/001/2
автомат-изированн-ый	09/02/103/01
автомат-ически-+	06/ 01/152/01
автомат-изованн-ый	07/02/103/01
автомат-ическ-ий	05/ 02/106/01
……	……..

Фрагмент словаря СОС приведен в табл. 14. Словарь представлен в прямом лексикографическом порядке. В каждой словоформе через знак *дефис* выделена словообразовательная основа, суффикс (сочетание суффиксов) и грамматическое окончание. При этом каждая словоформа словаря также сопровождается информацией о длине суффикса (сочетаний суффиксов), длине грамматического окончания, номера флективного класса и модели управления.

Порядок обработки слов алгоритмом словообразовательного анализа выполняется следующим образом. Вначале в составе слова устанавливается словоизменительная основа, грамматическое окончание и флективный класс слова. Эта задача решается путем использования ранее рассмотренного словоизменительного анализа (см. Алгоритм 1). Далее производится поиск на полное совпадение словоизменительной основы анализируемого слова и одной из словоизменительных основ словаря СОО. При положительном совпадении информация словоформы словаря СОО переносится на анализируемое слово. При несовпадении указанных словоизменительных основ необходимо будет выполнить полный цикл словообразовательного анализа. В этом случае на основе информации о словоизменительной основе и номере ее флективного класса необходимо определить в составе анализируемого слова словообразовательную основу и суффикс или их сочетание. Для этого нужно соотнести конечное буквосочетание словоизменительной основы с суффиксом (сочетанием суффиксов) максимальной длины. В случае их удачного соотнесения необходимо проверить по словарю СОО имеется ли в нем такая словообразовательная основа. Если проверка успешно выполнена, то в анализируемом слове правильно выделена словообразовательная основа и суффикс (сочетание суффиксов). В случае отсутствия в словаре такой основы, производится ее приближенное выделение с проверкой на совместимость выделенной основы и суффикса или их сочетаний.

Приведем алгоритм морфологического словообразовательного анализа русских слов [3].

Алгоритм 3 (алгоритм морфологического словообразовательного анализа русских слов).

Шаг 1. Выполняется обработка анализируемого слова процедурой морфологического словоизменительного анализа. На основе результатов этой обработки определяется словоизменительная основа, флективный класс слова и модель управления слова. Переход к шагу 2.

Шаг 2. Выполняется поиск словоизменительной основы анализируемого слова с установленным флективным классом в словаре СОО. В случае успешного поиска такой основы слову назначается ее словообразовательная грамматическая информация и выполняется переход к шагу 7. В случае неудачного поиска в словаре – переход к шагу 3.

Шаг 3. Выполняется последовательное соотнесение конечного буквосочетания анализируемого слова с суффиксами (сочетаниями суффиксов) словаря СоКС_И. В случае успешного соотнесения конечного буквосочетания переход к шагу 4. При отсутствии совпадения длина суффикса устанавливается равной нулевому значению. Переход к шагу 6.

Шаг 4. Начальная часть словоизменительной основы анализируемого слова (без совпавшего конечного буквосочетания) сравнивается на полное совпадение со словообразовательными основами словоформ словаря СОС. В случае успешного сравнения начального буквосочетания анализируемого слова с одной из словообразовательных основ словаря СОС вычисляется длина суффикса (сочетания суффиксов). Переход к шагу 6. При неудачном сравнении переход к шагу 5.

Таблица 15

Результаты работы морфологического словообразовательного анализа

Буквенный код исходного слова	Символ класса	Грамматич. характеристики слова: длина суфф. и оконч., флект. класс, мод. упр, род, число, падеж, лицо
российская	*A*	*03/02/106/01/2110*
сторона	*N*	*00/01/056/01/2110*
поставила	*L*	*02/01/125/34/2100*
вопрос	*N*	*00/00/001/01/1110,1140*
о	*F*	*00/00/164/46/0400,0600*
неполном	*A*	*00/02/103/01/1160,3160*
исполнении	*N*	*03/01/073/02/3130*
резолюции	*N*	*00/01/061/01/2120,2130,2160,2210,2240*

Шаг 5. Производится приближенное установление псевдоосновы и суффикса (сочетания суффиксов). В случае успешного установления псевдоосновы вычисляется длина суффикса. В случае невозможности ее установления длина суффикса устанавливается равной нулевому значению. Переход к шагу 6.

Шаг 6. Слову назначается словоизменительная информация и дополнительно приписывается к словообразовательной информации вычисленное значение длины суффикса (сочетаний суффиксов). Переход к шагу 7.

Шаг 7. Преобразование полученных результатов в структуру метаданных.

В табл. 15 приводятся результаты работы морфологического анализа русских слов.

В этой таблице используются мнемонические обозначения классов слов из табл.1, а значения цифр четырехзначных представлений грамматических признаков из табл. 7.

МОРФОЛОГИЧЕСКИЙ СЛОВООБРАЗОВАТЕЛЬНЫЙ СИНТЕЗ

Морфологический словообразовательный синтез – это лингвистическая процедура, реализующая генерацию требуемой формы слова словообразовательной парадигмы слова на основе информации об исходной форме слова и заданного набора грамматической информации. В общем виде словообразовательный синтез можно представить как процесс определения словообразовательной основы слова и последующего присоединения к ней требуемых суффиксов или их сочетаний с соответствующим грамматическим окончанием. При этом необходимо отметить, что при словообразовательном синтезе надо разрешить все проблемы, относящиеся к проблемам словоизменительного синтеза, а также решить основную проблему словообразовательного синтеза – получение достоверной информации о совместимости словообразовательной основы и присоединяемого к ней суффикса генерируемой формы слова [9].

Проверку совместимости словообразовательной основы слова и присоединяемых к ней новых суффиксов можно было бы проводить с помощью словаря СОС. Но для этого необходимо иметь всеобъемлющий словарь словообразовательных основ слов, в котором для каждой основы должен быть указан номер ее словообразовательного класса или (в случае омонимии основ) сочетание номеров классов. Поскольку словарь СОС имеет ограниченный размер и поэтому могут встречаться слова с «новыми» словообразовательными основами, то этим методом можно проверить только случаи содержащихся в словаре этих вновь образованных словоизментительных (слообразовательная основа и присоединенный к ней суффикс) основ в имеющемся словаре СОС путем поиска совпавших словоизменительных основ слов словаря. Если сформированная словоизменительная основа содержится в словаре словоизменительных основ, то это означает, что она правильная, а словообразовательная основа и присоединенный к ней суффикс (сочетание суффиксов) совместимы. Если не содержится, то эта словоизменительная основа вероятно либо неправильная, либо в словаре СОС такая словоизменительная основа отсутствует.

Для таких случаев использовался второй метод проверки, базирующийся на следующей гипотезе: *если два суффикса или их сочетания входят в один или несколько словообразовательных классов, то считается, что эти суффиксы или их сочетания с большой вероятностью принадлежат одной словообразовательной парадигме и словообразовательная основа синтезируемого слова, и присоединяемый суффикс (сочетание суффиксов) совместимы*. На основе совместного применения этих двух методов был построен, приведенный ниже, алгоритм морфологического словообразовательного синтеза русских слов.

Алгоритм 4 (алгоритм морфологического словообразовательного синтеза русских слов).

Шаг 1. Выполняется обработка анализируемого слова процедурой морфологического словообразовательного анализа (см. Алгоритм 3). В результате такой обработки определяется длина суффиксов (сочетания суффиксов) и длина грамматического окончания, флективный класс слова, а также набор его грамматических признаков. Если длина суффиксов (сочетания суффиксов) имеет нулевое значение, то переход к шагу 6. Если ненулевое – то переход к шагу 2.

Шаг 2. Выполняется отделение суффикса или сочетания суффиксов от словообразовательной основы анализируемого слова и присоединение к ней синтезирующего суффикса (сочетания суффиксов). Переход к шагу 3.

Таблица 16

Результаты работы морфологического словообразовательного анализа

Буквенный код исходного слова	Символ класса исх. слова	Грамматическая информация о слове	Грамматическая информация для синтеза слова	Буквенный код синтезируемого слова
российская	*A*	*03/02/106/2110*	*N/2150*	*россией*
сторона	*N*	*00/01/056/2110*	*N/2110*	*стороной*
поставила	*L*	*02/01/125/2100*	*A/1120*	*поставленного*
вопрос	*N*	*00/001/1110,1140*	*N/1250*	*вопросами*
о	*F*	*00/00/164/0400,0600*	*F/0000*	*о*
неполном	*A*	*01/02/103/1160,3160*	*K/2100*	*неполна*
исполнении	*N*	*03/02/106/2110*	*I/0000*	*исполнить*
резолюции	*N*	*00/01/073/3130*	*N/2250*	*резолюциями*

Шаг 3. Выполняется поиск в словаре СОС вновь образованной основы со словоизменительными основами словаря. В случае успешного сопоставления этих основ переход к шагу 6. При неудачном сравнении переход к шагу 4.

Шаг 4. Осуществляется проверка по словарю СоКС_И на принадлежность суффиксов (сочетаний суффиксов) анализируемого слова и словоформы с совпавшей словообразовательной основой словаря СОС одному или нескольким словообразовательных классам. В случае удачной проверки переход к шагу 6. При неудачной проверке – переход к шагу 5.

Шаг 5. При неудачной проверке устанавливается нулевое значение длины суффикса. Переход к шагу 6.

Шаг 6. К полученной словоизменительной основе присоединяется синтезирующее грамматическое окончание. Переход к шагу 7.

Шаг 7. Преобразование полученных результатов в структуру метаданных.

Результаты работы морфологического анализа русских слов приводятся в табл. 16.

В табл. 16 используются мнемонические обозначения классов слов из табл. 1, а значения цифр четырехзначных представлений грамматических признаков – табл. 7.

НОРМАЛИЗАЦИЯ РУССКИХ СЛОВ

Нормализация русских слов – это лингвистическая процедура, предназначенная для отожествления множества различных форм слов словоизменительной или словообразовательной парадигм путем их сведения в одну нормализованную каноническую форму слова [10]. При этом процедуру отождествления слов можно свести к отождествлению их канонических форм. Но при отождествлении слов возможен и другой подход, заключающийся в том, что текстовая форма анализируемого слова заменяется средствами морфологического синтеза на множество эквивалентных ей по смыслу нормализованных словообразовательных вариантов, а текстовая форма другого слова нормализуется только на уровне словоизменения. Затем эта форма второго слова сравнивается со всеми нормализованными словообразовательными вариантами первого и в случае совпадения с одним из них она считается эквивалентной по смыслу первому слову.

Обычно под нормализованной (канонической) формой слова понимается такая форма, которая традиционно указывается в словарях. Например, для существительного это форма именительного падежа единственного или (в случае *pluralia tantum*) множественного числа, для глагола – форма инфинитива, для прилагательного – форма именительного падежа единственного числа мужского рода.

Как уже выше было сказано, в рамках используемой теоретической концепции необходимо различать два уровня нормализации: на уровне словоизменения и на уровне словообразования. При нормализации слов на словоизменительном уровне каноническая форма слова должна представлять всю его словоизменительную парадигму. Например, в словоизменительной парадигме: *телефон, телефона телефону, телефоном, телефоне, телефоны, телефонов, телефонам, телефонами, телефонах – в качестве канонической формы* можно представить члена этой парадигмы *«телефон»*). При нормализации слов на словообразовательном уровне каноническая форма слова должна представлять по возможности всю его словообразовательную парадигму. Например, тождественные или близкие по смыслу формы слов – *испытают, испытав, испытавший, испытай, испытайте, испытал, испытан, испытание, испытанный, испытать, испытывавший, испытываемый, испытывают, испытывай, испытывайте, испытывать, испытывал* принадлежащие к различным частям речи могут быть заменены на одну форму отглагольного существительного «*испытание*». А формы слов *замолчит замолчав замолчавши замолчавший замолчал замолчать замолчи замолчите* — на одну форму инфинитива «*замолчать*» (в составе словообразовательной парадигмы нет отглагольного существительного).

Выбор канонической формы слова, представляющей множество его словообразовательных вариантов и имеющих примерно одинаковый смысл, должен производиться с учетом системы словообразования русского языка. По аналогии с канонической формой слова представляющей множество его словоизменительных

Таблица 17

Результаты работы процедуры нормализации русских слов

Буквенный код исходного слова	Символ класса исх. слова	Грамматическая информация о слове	Буквенный код нормализованного слова	Символ класса норм. слова
Нормализация на уровне словоизменения				
российская	*A*	*02/106/2110*	*российский*	*A*
сторона	*N*	*01/056/2110*	*сторона*	*N*
поставила	*L*	*01/125/2100*	*поставил*	*L*
вопрос	*N*	*00/001/1110,1140*	*вопрос*	*N*
о	*F*	*00/164/0400,0600*	*о*	*F*
неполном	*A*	*02/103/1160,3160*	*неполный*	*A*
исполнении	*N*	*01/073/3130*	*исполнение*	*N*
Нормализация на уровне словообразования				
российская	*A*	*03/02/106/2110*	*россия*	*N*
сторона	*N*	*00/01/056/2110*	*сторона*	*N*
поставила	*L*	*02/01/125/2100*	*поставить*	*I*
вопрос	*N*	*00/001/1110,1140*	*вопрос*	*N*
о	*F*	*00/00/164/0400,0600*	*о*	*F*
неполном	*A*	*01/02/103/1160,3160*	*неполный*	*A*
исполнении	*N*	*03/02/106/2110*	*исполнение*	*N*
резолюции	*N*	*00/01/073/3130*	*резолюция*	*N*

вариантов для представления множества членов словообразовательной парадигмы можно также ввести свою каноническую форму. Здесь можно предложить следующее правило: *в качестве каноническую формы словообразовательной парадигмы могут выступать следующие классы слов в порядке уменьшения их приоритетов : существительное, если оно является членом парадигмы или (если нет существительного), то инфинитив. В тех случаях, когда в составе парадигмы нет ни существительного, ни инфинитива, то в качестве канонической формы может выступать прилагательное. Если и прилагательного нет, то любая другая заранее заданная форма* [9].

Переход от вариантной словообразовательной формы слова к его канонической можно представить себе как замену суффикса или сочетания суффиксов вариантной формы на суффикс (сочетание суффиксов) канонической формы. Для этого необходимо уметь выделять в слове его словообразовательную основу и суффиксы иметь ассоциативный словарь суффиксов, в котором для каждого суффикса (сочетания суффиксов) будет указан один или несколько вариантов его замены на суффикс или на сочетание суффиксов соответствующей канонической формы и иметь процедуру проверки правильности такой замены (проверки совместимости словообразовательной основы слова и присоединенных с ней суффикса или сочетания суффиксов).

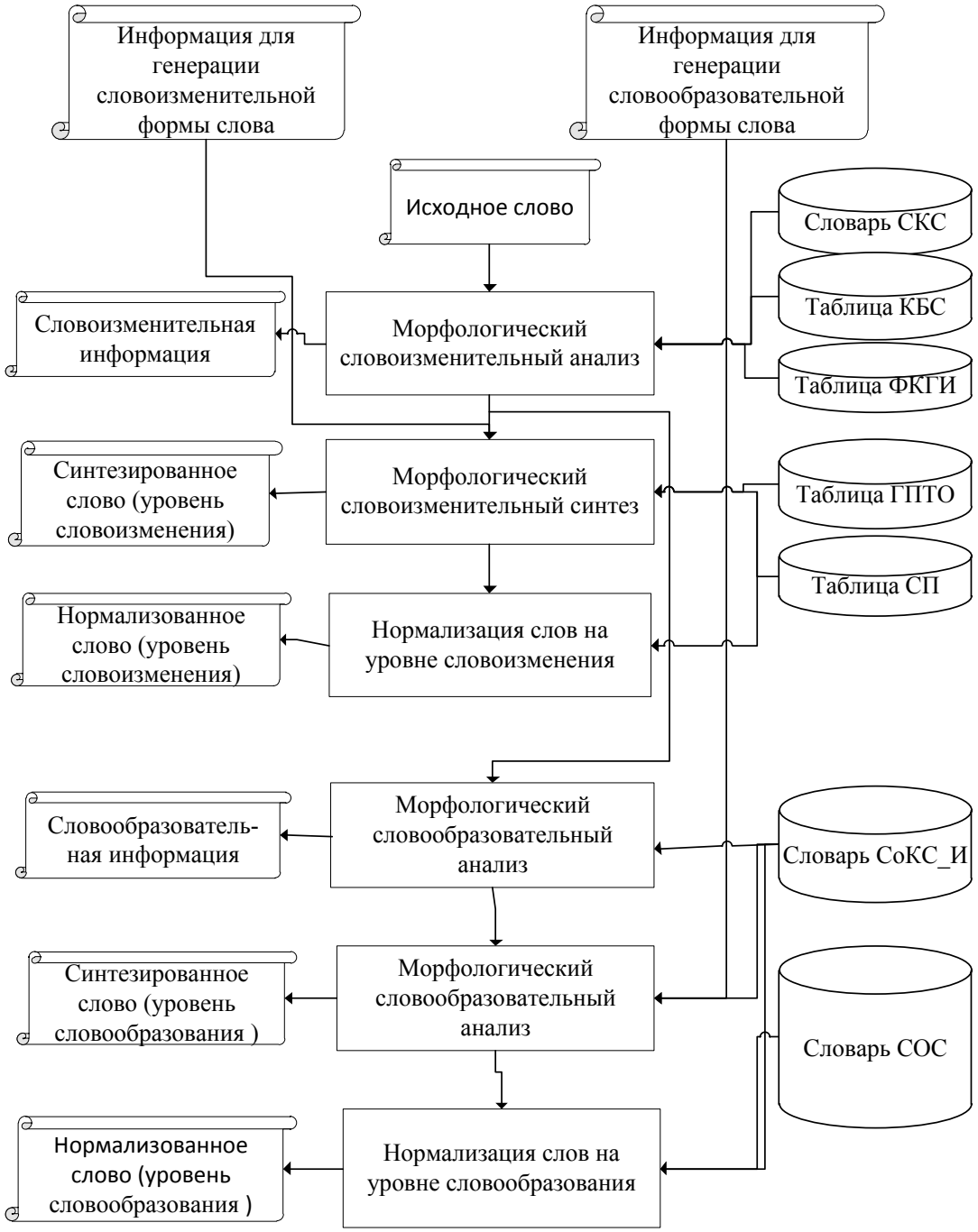
Процедуру нормализации нужно рассматривать как реализацию частного случая процедур словоизменительного или словообразовательного синтеза, когда для каждого лексико-грамматического класса заранее задается заранее заданная нормальная форма слова. Процедуре нормализации обязательно должна предшествовать процедура морфологического анализа, в результате которой определяется структура слова и назначается набор грамматических признаков. В процессе нормализации слов на словоизменительном уровне производится, как правило, замена текстового окончания на нормализующее, при этом, в некоторых случаях требуется трансформация конечного буквосочетания основы слова. На словообразовательном уровне производится словообразовательная трансформация вариантной формы слова в каноническую. Результаты работы процедуры нормализации русских слов приведены в табл. 17.

В табл. 17 используются мнемонические обозначения классов слов из табл. 1, а значения цифр четырехзначных представлений грамматических признаков – табл. 7.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МАШИННОЙ ГРАММАТИКИ РУССКОГО ЯЗЫКА

На базе разработанных алгоритмов и декларативных средств был создан программный модуль машинной грамматики русского языка. Общая архитектурная схема модуля представлена на рисунке.

Анализ функционирования этого модуля в промышленном режиме показал, что словоизменительный анализ и синтез имеет высокое быстродействие, а вероятность правильного назначения информации составляет 99,5%. Словообразовательный анализ и синтез показывает на порядок меньшее быстродействие, а вероятность правильного назначения информации не превышает 88%. Автоматическая нормализации слов на уровне словоизменения и словообразования по быстродействию и точности назначения грамматической информации соответствует быстродействию и точности процедурам словоизменительного и словообразовательного анализа.



Общая схема информационно-технологической архитектуры модуля машинной грамматики русского языка

ЗАКЛЮЧЕНИЕ

Подводя итоги рассмотрения процесса создания программных и декларативных средств машинной грамматики, мы хотели бы еще раз подчеркнуть важность принципов, методов и средств, положенных в основу формализации грамматики русского языка. К ним можно отнести:

1. Таблица флективных классов, разработанная проф. Г.Г. Белоноговым в конце 60-ых годов прошлого века. Эта таблица позволила представить все многообразие словоизменения русского языка в компактном виде. В основу этой таблицы были положены основные типы словоизменения русских слов [3].

2. Словарь словообразовательных классов русского языка, включающей в структурированном виде все основные трансформации русских слов [5].

3. Принцип лингвистической аналогии, который позволил выявить и реализовать трансформационные закономерности словоизменения и словообразования, многократно сократить объемы словарей и грамматических таблиц, а также успешно решить задачи, не поддающиеся решению алгоритмическими методами [11].

4. Созданный комплекс декларативных средств машинной грамматики полностью покрывает все возможные реализации по автоматическому определению структуры слов, установлению их грамматических характеристик, а также средств установления

смыслового тождества между членами одной словоизменительной и словообразовательной парадигмы.

5. Широкое использование средств автоматизации при формировании декларативных средств позволили многократно сократить трудозатраты и существенно повысить качество созданных словарных ресурсов.

6. Разработанные программные средства базируются на уникальных алгоритмах, обеспечивающих их быстродействие и высокое качество обработки текстовой информации.

Описанные программные и декларативные средства машинной грамматики русского языка в настоящее время широко используются для решения достаточно сложных задач автоматической обработки и анализа смыслового содержания текстовой информации в ряде промышленных информационных систем, таких, например, как системы машинного перевода, поисково-аналитические системы, системы обеспечения информационной поддержки жизненного цикла сложных инженерных объектов, системы управления знаниями высокотехнологичных отраслей и др. [12].

СПИСОК ЛИТЕРАТУРЫ

1. Мельчук И.А. Курс общей морфологии. Том 1. – Москва-Вена: «Языки русской культуры» Венский славистический альманах, Издательская группа «Прогресс», 1997.
2. Зализняк А.А. Грамматический словарь русского языка. – М.: «Русский язык», 1977.
3. Белоногов Г.Г., Калинин Ю.П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Компьютерная лингвистика и перспективные информационные технологии // Научно-техническая информация. Сер.2. – 2004. – № 8. – С. 22-32.
4. Белоногов Г.Г., Калинин Ю.П., Поздняк М.В., Хорошилов А.А. Алгоритм многоступенчатого морфологического анализа русских слов // Научно-техническая информация. Сер.2.– 1983. – № 7. – С. 6-10.
5. Белоногов Г.Г., Кузнецов Б.А., Новоселов А.П., Хорошилов А.А. и др. Приложение 2. Словообразовательные классы слов Автоматизированная обработка научно-технической информации // Итоги науки и техники. Серия «Информатика». – М.: ВИНИТИ, 1984. – С. 163-296.
6. Белоногов Г.Г., Калинин Ю.П., Поздняк М.В., Хорошилов А.А., Яфаева Г.М. Инвертированный словарь словообразовательных классов слов. – М.: ВИНИТИ, 1983. – 88 с. – Деп. в ВИНИТИ 10.05.83, № 2502-83.
7. Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов А.А. Методы автоматизированного составления и ведения словарей // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – №3. – С. 91–97.
8. Белоногов Г.Г., Зеленков Ю.Г., Кузнецов Б.А., Новоселов А.П., Панова Н.С., Рыжова Е.Ю., Хорошилов А.А., Штурман Я.П. Машинный политематический словарь основ слов // Научно-техническая информация. Сер.2. – 1988. – № 9. – С. 26-29.
9. Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А., Хорошилов-мл. А.А. Автоматическое распознавание смыслового тождества и смысловой близости русских слов на основе их словообразовательного анализа и синтеза // Научно-техническая информация. Сер. 2. – 2003. – №1. – С. 30–33.
10. Новоселов А.П., Хорошилов А.А., Хорошилов А.А. и др. Алгоритм автоматической нормализации слов // Вопросы информационной теории и практики. – 1985. – № 53. – С. 67-71.
11. Белоногов Г.Г., Зеленков Ю.Г., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Метод аналогии в компьютерной лингвистике // Научно-техническая информация. Сер.2. – 2000. – № 1. – С.21-31.
12. Белоногов Г.Г., Гиляревский Р.С., Селедков С.Н., Хорошилов А.А. О путях повышения качества поиска текстовой информации в системе Интернет // Научно-техническая информация. Сер. 2. – 2013. – №8. – С. 15–22; Belonogov G.G., Gilyarevskii R.S., Seletkov S.N., Khoroshilov A.A. Ways to Improve the Quality of Textual Data Searches on the Internet // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 47, № 4. – P. 111-120.

Материал поступил в редакцию 18.04.18.

Сведения об авторах

АБЛОВ Игорь Васильевич – начальник управления 27-го Центрального научно-исследовательского института Министерства обороны России, Москва
e-mail: iablov@mail.ru

КОЗИЧЕВ Вячеслав Николаевич – доктор технических наук, главный научный сотрудник 27-го Центрального научно-исследовательского института Министерства обороны России, Москва
e-mail: vkozichev@mail.ru

ШИРМАНОВ Александр Викторович – кандидат технических наук, заместитель начальника 27-го Центрального научно-исследовательского института Министерства обороны России, Москва
e-mail: avshirmanov@mail.ru

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор МАИ, ведущий научный сотрудник Федерального исследовательского центра «Информация и управления» Российской академии наук, Москва
e-mail: khoroshilov@mail.ru

ХОРОШИЛОВ Алексей Александрович – кандидат технических наук, научный сотрудник Федерального исследовательского центра «Информация и управления» Российской академии наук, Москва
e-mail: ahoroshilov@mail.ru

НАМ ПИШУТ

УДК 001.894:81’322.4

М.М. Гольдреер

Адаптивный контекстно-тематический машинный перевод

Описано изобретение, относящееся к области машинного перевода текста на естественном языке при непосредственном общении пользователей, техническим результатом которого является повышение точности машинного перевода с языка пользователя на иностранный язык, а также универсализации и экономии вычислительных мощностей. В систему вводят текст на языке пользователя отдельными исходными предложениями, для которых она подбирает аналогичные фразы на иностранном языке, хранящиеся в ее базе данных, по определенным темам. Новые темы и стандартные фразы к ним точными переводами встраивают в систему по мере необходимости.

Ключевые слова: *машинный перевод, непосредственное общение пользователей, переводческая память*

В настоящее время растет рынок машинного перевода с широким применением цифровых интернет-технологий. Однако, несмотря на их бурное развитие, качество такого перевода практически не улучшается, все усилия сводятся к тому, чтобы по возможности точнее передать общий смысл переводимых текстов, но точность в этом случае относительна. Машинный перевод может сэкономить время специалистам при ознакомлении со специализированными текстами, но практически непригоден в межличностном общении разноязыких людей на бытовые темы при обычном разговоре. В то же время уже есть программные приложения, позволяющие делать машинный перевод не только текстовых, но и голосовых сообщений. В целом все разрабатываемые методики и системы современного машинного перевода имеют конечную цель – воспроизвести как можно точнее работу переводчика-человека, специалиста самого высокого класса. Для этого создаются сложнейшие системы структурного и статистического анализа текстов на различных языках, задействуются все более мощные вычислительные ресурсы. Но решение задачи не просматривается, ибо это задача создания полноценного искусственного интеллекта.

Варианты воплощения настоящего изобретения относятся к реализуемому при помощи компьютера способу предоставления информации автоматической системе машинного перевода для повышения точности перевода с языка пользователя на иностранный язык, его универсальности, а также для экономии вычислительных мощностей. Это включает прием исходного текста на языке пользователя и подачу его отдельными предложениями-фразами для перевода. Автоматическая система машинного перевода ищет в своих базах данных стандартные фразы – аналоги полученным исходным фразам и темы, которым соответствуют найденные фразы-аналоги. После отбора пользователем нужной темы, если стандартная фраза-аналог исходной фразы представлена в нескольких темах в базе данных системы, автоматическая система машинного перевода дает этой фразе однозначно точный перевод на иностранные языки, которые заготовлены в ее базе данных. Если же стандартная фраза представлена только в одной теме, то она переводится сразу, без подбора темы пользователем. Если какие-то фразы из исходного текста не имеют стандартных фраз-аналогов и своих тем в автоматической системе машинного перевода, то в процессе перевода исходного текста эти новые фразы-аналоги создаются и закладываются в базы данных автоматической системы машинного перевода через связанный с ней удаленный источник надежной информации.

Наиболее близким аналогом предлагаемого изобретения является «Адаптивный машинный перевод», патент RU 2382399 C2 от 18.06.2004 г. Недостаток этого аналога в том, что он ориентирован на перевод сразу всего заданного текста, подвергая его сложному грамматическому, семантическому и экстралингвистическому анализу, после чего полученный перевод подвергается статистическому сравнению с похожими текстами и дорабатывается уже после этого сравнения. Но поскольку любой язык постоянно и быстро меняется, вбирая в себя новые слова, смыслы, подтексты и термины, то такая методика перевода обречена на неизбежное отставание и использование

все больших вычислительных мощностей, никогда не выходя за понимание общего смысла переводимого текста и не давая исчерпывающего представления о его деталях и тонкостях. А для чисто разговорного общения между людьми этот способ вообще непригоден, так как может менять в течение непродолжительного времени перевод одних и тех же фраз.

Предлагаемое же нами изобретение упрощает задачу для автоматической системы машинного перевода тем, что пользователь приспосабливается к ее ограниченным возможностям, задавая для перевода не произвольные тексты целиком, а отдельные фразы, каждая из которых представляет законченный мини-контекст для каждого своего слова и входит в группу фраз, соответствующих определенной теме, которая тоже задается в качестве команды. Таким образом, автоматическая система машинного перевода просто ищет в заданной теме стандартные фразы – аналоги фразе, заданной пользователем, и, если выбран нужный аналог, то пользователь тут же получает перевод в самом точном и однозначном виде из базы данных автоматической системы.

Если же какой-то темы с соответствующими ей стандартными фразами и их переводами еще нет в системе машинного перевода, то с помощью источника надежной информации, в котором могут быть задействованы и переводчики-специалисты, всегда может быть создана новая тема и наполнена соответствующими стандартными фразами с их точными переводами и постоянно пополняемыми наборами фраз – нестандартных соответствий. Таким образом, экономится время пользователей, особенно при личном диалоговом общении, и нет необходимости в больших вычислительных мощностях для сложного анализа задаваемых текстов. Тем же, кто хочет перевести иностранный текст на свой язык, эта методика пока не поможет, зато она может резко облегчить и ускорить работу профессиональных переводчиков-специалистов, которые, зная иностранный язык и переводя на него тексты, могли бы к тому же пополнять базы данных системы автоматического машинного перевода созданными ими стандартными парами «фраза – перевод» с соответствующими новыми темами.

Предлагаемый нами способ значительно облегчает перевод исходных текстов на иностранные языки, так как уже имея один готовый перевод исходного текста, можно этот перевод задавать системе, и она его легче переведет на другие языки, поскольку он будет состоять из одних стандартных фраз с уже выбранными для них темами. Представляется, что по сравнению с уже имеющимися системами автоматического машинного перевода, с тем же номером патента 2382399, этот способ упрощает процесс перевода и своей адаптации за счет того, что весь анализ переводимых фраз сводится к поиску аналогов в базах данных путем сравнения, а адаптация (обучение!) – к пополнению баз данных новыми темами с наборами новых стандартных пар «фраза – перевод» с сопутствующими им наборами нестандартных фраз-аналогов.

Однако это не исключает в будущем встраивания и использования в описываемом изобретении уже имеющихся методик компьютерного анализа и перевода, а также вероятностных и статистических методов, которые применяются в способе машинного перевода по патенту 2382399 для повышения степени автоматизации процесса перевода. Но и в этом случае задача будет резко упрощена, поскольку работать придется с отдельными фразами без ориентации на смысл и специализацию всего массива похожих текстов, как это происходит сейчас. Таким образом достигается универсальность описываемого изобретения для максимально точного перевода любых, а не только близких по тематике, текстов как в уже действующих системах автоматического машинного перевода, а общее упрощение работы системы может многократно снизить требования к мощности вычислительных ресурсов для ее обеспечения.

Представленный способ машинного перевода можно использовать с множеством любых вычислительных систем, сред или конфигураций вычислительных систем общего или специального назначения. Хорошо известные подобные системы включают персональные компьютеры, серверы, карманные или портативные устройства, мультипроцессорные системы, системы на основе микропроцессоров, приставки, программируемую бытовую электронику, сетевые ПК, мини-компьютеры, универсальные ЭВМ, телефонные системы, распределенные вычислительные среды, содержащие любые из перечисленных систем или устройств и т.д.

Материал поступил в редакцию 01.02.18

Сведения об авторе

ГОЛЬДРЕЕР Михаил Маркович – изобретатель, Волгоградская обл., г. Волжский
e-mail: mg@rucosm.com