ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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. **№ 4** **Москва 2015**

ОБЩИЙ РАЗДЕЛ

УДК [002 : 004] : 001.89

Н.Е. Каленов

Об информационном сопровождении фундаментальных научных исследований

Рассмотрены особенности библиотечной деятельности в новых условиях формирования и хранения библиотечных ресурсов. Приведены аргументы, обосновывающие важную роль библиотек в процессах информационного сопровождения научных исследований. Показана структура библиотечной сети Федерального агентства научных организаций (ФАНО) – ранее Российской академии наук (РАН). На примере БЕН РАН рассмотрены задачи Центральной библиотеки сети в качестве традиционной библиотеки, информационного центра, научно-исследовательского института, технического центра.

Ключевые слова: *информационное обеспечение научных исследований, библиотечная деятельность, печатные издания, электронные ресурсы, библиотечная сеть РАН и БЕН РАН, задачи библиотеки как информационного центра*

Научная информация, под которой в данном контексте мы понимаем опубликованные в любой форме и на любом материальном носителе документы, отражающие полученные знания, является неотъемлемой частью любых (как фундаментальных, так и прикладных) исследований. Этот тезис является аксиомой и не требует дополнительных объяснений. Научная информация играет двоякую роль в фундаментальных исследованиях. С одной стороны, она является основным источником развития научной мысли (не зная текущего состояния науки, нельзя ее развивать), с другой стороны, опубликованные мате-

риалы являются основным результатом во многих областях фундаментальной науки, и их хранение (неважно, на каких носителях) необходимо как свидетельство успехов, достигнутых цивилизацией.

Таким образом, в процессе развития общества должно быть обеспечено: а) распространение научной информации и б) хранение научной информации.

До недавнего времени не вызывал сомнения тезис о том, что для распространения научной информации должна существовать отдельная служба – связующее звено между научным коллективом и информационным пространством, обеспечивающее информирование и предоставление исследователям необходимых им ресурсов. Хранение информации также логично было поручить этой службе, исходя из требований к оперативности предоставления требуемой информации. В качестве такой службы во всех областях фундаментальных исследований вплоть до середины 20-го века выступали научные библиотеки. Они собирали, хранили и предоставляли ученым необходимую им информацию. Развитие научных библиотек в мире, в частности, в России шло одновременно с развитием научной инфраструктуры. Бурное развитие науки в XX веке, формирование мощной сети научных учреждений и, как следствие, лавинообразный поток научной информации обусловили необходимость целенаправленного информирования ученых о появлении в мире новых результатов в исследуемых ими областях науки. Во многих странах стали создаваться национальные и отраслевые информационные центры – генераторы вторичной информации. В Советском Союзе это были национальные центры: Всероссийский институт научной и технической информации (ВИНИТИ) РАН, охватывающий тематику точных, естественных и технических наук, Институт научной информации по общественным наукам (ИНИОН) РАН, охватывающий тематику по общественным наукам), Всероссийский научно-исследовательский институт медицинской информации (ВНИИМИ), специализирующийся на обработке информации по медицине и др. Функции доведения информации до конкретных пользователей выполняли как научные библиотеки (в первую очередь, академические), так и сеть отраслевых центров научно-технической информации (НТИ), возглавляемая организацией Росинформресурс, большинство из которых имели в своем составе научные библиотеки.

В последней четверти XX века центральные академические библиотеки фактически превратились в информационно-библиотечные центры. В частности, Библиотека по естественным наукам (БЕН) создавалась в 1973 г. как информационно-библиотечный центр, обеспечивающий научной информацией исследовательские организации центральной части Академии наук на основе современных информационных технологий. Начиная с конца 1970 гг., БЕН информировала ученых о новых публикациях по тематике их исследований, предоставляла копии статей по заказам ученых, оптимизировала приобретение информационных ресурсов и т.п.

В настоящее время БЕН РАН, Библиотека Академии наук (БАН) и центральные библиотеки региональных отделений РАН (Центральная научная библиотека Уральского отделения РАН – ЦНБ УрО РАН, Государственная публичная научно-техническая библиотека – ГПНТБ СО РАН, Центральная научная библиотека Дальневосточного отделения – ЦНБ ДВО РАН) являются мощными информационными центрами, обеспечивающими сопровождение научных исследований в академических организациях на основе современных сетевых технологий.

В последние годы в обществе и, в том числе, в научной среде стало формироваться мнение, что «в Интернете все есть», поэтому ученый сам найдет в сети то, что его интересует, информационно-библиотечные услуги ему не требуются, а научные библиотеки больше никому не нужны. Печатные издания надо перевести в электронную форму, после чего перестать их (печатные издания) хранить.

Такой подход свидетельствует о поверхностном отношении к процессам информационного сопровождения научных исследований и при достаточно серьезном анализе не выдерживает критики. Приведем следующие аргументы.

1. Лавинообразное нарастание объемов научной информации (в том числе, представленной в Интернете), взаимопроникновение различных научных областей обусловливают значительные временные затраты на поиск и отбор действительно нужных ресурсов. Для их уменьшения, повышения точности и полноты поиска информации требуются специальные навыки, которыми большинство ученых не обладают. Если каждый исследователь будет сначала осваивать методы поиска информации, а затем самостоятельно отслеживать появление в мире новых данных в интересующей его области науки, то у него не останется времени на проведение собственно исследований. Поэтому логично, чтобы работу по анализу мирового научного информационного пространства и отбору материалов, представляющих интерес для исследователей, занимающихся определенной проблемой, выполняли сотрудники, для которых информационный поиск является профессией. Они должны регулярно информировать ученых о появлении интересующих их ресурсов (это могут быть сетевые адреса ресурсов, библиография и рефераты статей по рассматриваемой проблеме, цитирование публикаций обслуживаемого коллектива ученых и т.п.).

2. Серьезные научные ресурсы, даже если они и представлены в Интернете, стоят достаточно дорого. Большинство ведущих зарубежных издательств и научных обществ, таких как Elsevier, Springer, American Physical Society, American Chemical Society и др., предоставляют в свободном доступе только библиографическую информацию (описание статей) и (в лучшем случае) рефераты. Организация сетевого доступа к коммерческим источникам научной информации требует значительной по объему и сложности специфической работы, связанной с выбором нужных ресурсов, проведением переговоров с поставщиками, согласованием условий предоставления ресурсов, заключением контрактов и оформлением лицензионных соглашений, предоставлением IP-адресов и контролем выполнения договорных обязательств.

Для научных сотрудников такая деятельность не является характерной, поэтому сложившаяся мировая практика организации сетевого доступа к научной информации состоит в том, что ею занимаются библиотеки научных организаций и университетов. Библиотеки, в силу специфики своей деятельности, свободно «ориентируются» в информационном пространстве, имеют опыт взаимодействия как с издательствами и иными поставщиками научной информации, так и с потребителями информации; работа по информационному обеспечению научных исследований является их специальностью и прямой обязанностью. Кроме того, стремясь сократить затраты на приобретение доступа к сетевым ресурсам, научные библиотеки объединяются в консорциумы. Эта практика широко распространена в мире, и она требует также серьезной организационной и технической работы, требующей специальных навыков.

3. Полный отказ от печатных изданий и пользование исключительно сетевым доступом к научным журналам имеет ряд отрицательных сторон. Любые перебои связи, финансовые проблемы издающей организации или провайдера, переход коммерческого издательства к другому владельцу неизбежно вызывают перебои в получении информации. Получив печатное издание, библиотека становится его владельцем и в любой момент может предоставить его любому пользователю. Приобретая доступ к электронной версии журнала, пользователь не является собственником, и ему необходимо постоянно оплачивать поддержку доступа. Если в какой-то момент пользователь отказался от подписки на текущие выпуски журнала (изменилась тематика исследований, недостаточно финансовых ресурсов), то для сохранения доступа к его архивным выпускам (за которые было заплачено ранее как за текущие) необходимо заплатить поставщику определенную сумму. Это вполне естественно, поскольку поддержка доступа к ресурсам требует денег. Именно поэтому зарубежные библиотеки (выступающие в качестве посредников между учеными и информационным рынком) создают консорциумы по доступу к сетевым научным ресурсам (что позволяет значительно экономить финансовые ресурсы), но при этом приобретают «на всех» один-два «страховых» печатных экземпляра необходимых ученым журналов (проанализированный нами зарубежный опыт в этом направлении представлен в [1]).

4. Следующий фактор, определяющий необходимость специальной службы, обеспечивающей взаимодействие науки с информационным пространством, связан с проведением библиометрических исследований, результаты которых предлагается учитывать при оценке качества научных исследований. Такие исследования являются отдельной областью науки, требующей не только значительных временных затрат, но и специальной подготовки, без которой получение достоверных данных невозможно. Поэтому эти работы должны выполняться профессионалами, обладающими необходимыми навыками и (что достаточно важно) не заинтересованными в «субъективных» результатах. Такими профессионалами в настоящее время являются сотрудники центральных академических библиотек – БЕН РАН [2, 3], ГПНТБ СО РАН [4], ЦНБ УрО РАН [5] (мы привели эти публикации только для примера, они есть в каждой библиотеке).

5. Результаты деятельности научного коллектива должны отражаться в мировом информационном пространстве. В современных условиях это не только публикации в научных изданиях, но и многоаспектная информация, представленная в Интернете на сайте этого коллектива. Она должна включать базы данных публикаций сотрудников, полученных ими патентов, научных отчетов, докладов на конференциях, а также сведения о научных достижениях коллектива, полученных грантах и т.п. Очевидно, что отдельные исследователи не должны заниматься подобной работой, достаточно далекой от непосредственной научной деятельности. В то же время, эта работа, будучи поручена библиотечным сотрудникам, будет выполняться на высоком уровне, поскольку она полностью соответствует их профессиональным навыкам.

Как уже указывалось, основными подразделениями, обеспечивающими информационное сопровождение научных исследований, проводимых академическими организациями, являются библиотеки. Современная информационно-библиотечная деятельность – это специфическая, достаточно сложная область науки и технологий, успешно руководить которой могут специалисты именно в этой области. Поэтому в Академии наук эта деятельность строилась на принципах централизации, что предусматривает организационное взаимодействие библиотечных подразделений организаций РАН с целью:

- координации всех направлений информационно-библиотечной деятельности внутри Централизованной библиотечной системы (ЦБС), единого научно-методического руководства ею со стороны центральных библиотек;
- координированного взаимодействия с внешними организациями – библиотеками, издательствами, информационными и книготорговыми центрами;
- оптимизации расходования финансовых средств, выделяемых на приобретение и обработку информационных ресурсов;
- сбора и обработки отчетных данных для предоставления руководству РАН и центрам обработки библиотечной статистики.

При том, что библиотеки имеются в подавляющем большинстве академических институтов, они имеют различный статус – часть из них являются подразделениями институтов, часть – подразделениями центральных библиотек. В РАН (ныне в ФАНО) существуют 6 централизованных библиотечных систем – 4 региональные и 2 отраслевые. Каждая ЦБС возглавляется центральной библиотекой (ЦБ), имеющей статус научно-исследовательского института. ЦБС Санкт-Петербургского региона, возглавляемая Библиотекой Академии наук (БАН), включает 33 библиотеки, являющиеся отделениями ЦБ (их сотрудники входят в штат БАН, фонды числятся на балансе БАН), расположенные в 20-ти научно-исследовательских учреждениях (НИУ) РАН естественнонаучного профиля и 13-ти – гуманитарного. ЦБС Сибирского отделения РАН, возглавляемая ГПНТБ СО РАН

(г. Новосибирск), включает 67 библиотек, расположенных в Западной и Восточной Сибири, являющихся структурными подразделениями НИУ СО РАН, и одну, являющуюся отделением ЦБ в Академгородке. ЦБС Уральского отделения РАН, возглавляемая ЦНБ УрО РАН (г. Екатеринбург), включает 25 библиотек, 2 из которых являются отделениями ЦБ, остальные – подразделениями НИУ, относящимися к 6-ти научным центрам УрО РАН, расположенным в Уральском регионе, а также в Архангельске. ЦБС Дальневосточного отделения РАН, возглавляемая ЦНБ ДвО РАН (г. Владивосток), включает 19 библиотек, являющихся подразделениями Центральной библиотеки. Отраслевая ЦБС в области общественных и гуманитарных наук возглавляется Фундаментальной библиотекой, входящей в состав ИНИОН РАН, и включает 21 библиотеку, которые являются подразделениями ИНИОН в московских институтах гуманитарного профиля.

Наиболее крупной и сложной по структуре отраслевой ЦБС является система, возглавляемая Библиотекой по естественным наукам (БЕН) РАН. В ее состав входят 95 библиотек, 66 из которых входят в штатную структуру БЕН (причем некоторые библиотеки обслуживают несколько институтов). Два подразделения БЕН работают в подмосковных научных центрах – Пущино и Черноголовке. Поскольку в составе этих центров несколько НИИ, каждая из библиотек имеет свои филиалы в институтах центров, причем часть сотрудников филиалов входит в штат БЕН, часть – в штат соответствующих институтов. Общее количество таких «филиалов» – 15. Современное состояние ЦБС и предлагаемые формы их развития рассмотрим на примере ЦБС БЕН РАН.

Как традиционная библиотечная система ЦБС БЕН РАН имеет единый отраслевой фонд научной литературы, включающий около 12 млн ед. хранения. Эти фонды хранятся в Центральной библиотеке (около 1,5 млн экз.) и рассредоточены по отделам БЕН РАН в институтах (около 5 млн экз.) и по библиотекам – подразделениям институтов (более 5 млн экз.). БЕН, с момента своего образования (1973 г.), осуществляет централизованное приобретение и централизованную обработку всех изданий, поступающих в фонды ее ЦБС. До этого те же функции по отношению к тем же библиотекам осуществлял Сектор сети библиотек. В последние годы, наряду с отечественными изданиями, БЕН РАН централизованно обеспечивает доступ к зарубежным журналам, базам данных и коллекциям книг из библиотек своей ЦБС.

Весь фонд ЦБС БЕН РАН отражен в сводных карточных каталогах («замороженных» несколько лет назад) и электронных каталогах, доступных через Интернет (сводный электронный каталог журналов содержит данные о нескольких миллионах выпусков журналов, поступивших в ЦБС, начиная с 1990 г., сводный электронный каталог книг – данные о поступлениях, начиная с 1993 г.). Каждая библиотека, входящая в ЦБС БЕН РАН, ведет свои локальные каталоги, используя результаты централизованной обработки информации, осуществляемой специалистами ЦБ. Все каталоги поддерживаются в актуальном состоянии.

Библиотеки обслуживают ученых «своих» институтов на основе локальных фондов, а также по межбиблиотечному абонементу (МБА) через Центральную библиотеку, использующую как единый фонд ЦБС, так и фонды других библиотек страны, а также фонды зарубежных библиотек, направляя им заказы по международному МБА. Все технологические процессы в ЦБ и в ряде ее отделов в НИУ РАН автоматизированы, на сайте ЦБ представлены различные сервисы.

Как информационный центр БЕН РАН подключает к своим журнальным каталогам ссылки на полные тексты доступных ее пользователям журналов; в каталогах книг демонстрирует отсканированные информационные страницы; поддерживает на сайте ЦБ раздел «Естественные науки в Интернете», а на сайтах своих отделов – развернутую информацию по тематике обслуживаемых ими институтов; формирует на сайте по заказам институтов виртуальные выставки информации по заданным тематическим направлениям; ведет базы данных публикаций сотрудников институтов, а также формирует и поддерживает на своих серверах проблемно-ориентированные базы данных по различной тематике.

Как научно-исследовательский институт БЕН РАН ведет исследования в таких областях, как: применение современных методов информатики в технологических процессах информационно-библиотечного сопровождения научных исследований; создание и поддержка электронных библиотек; развитие систем классификации научной информации; управление информационными ресурсами, библиометрия. Технологические и программные решения, разработанные специалистами БЕН РАН, во многих случаях являются уникальными, не уступают, а часто опережают отечественные и зарубежные аналоги. Они внедряются не только в ЦБ и ее отделах, но и в других организациях. БЕН РАН участвует в разработке программного обеспечения и наполнении электронной библиотеки «Научное наследие России», постоянно получает гранты Российского фонда фундаментальных исследований (РФФИ) и Российского гуманитарного научного фонда (РГНФ), участвует совместно с другими организациями в выполнении НИР по государственным заказам по тематике своей научной деятельности.

Как технический центр БЕН РАН является одним из 8-ми узлов московского сегмента опорной телекоммуникационной сети [8]. Примерно на том же уровне работают центральные библиотеки региональных отделений РАН, акцентируя внимание на специфичных для своих регионов вопросах [6, 7].

Таким образом, организационные и технологические решения, реализованные в существующих ЦБС, могут служить хорошим фундаментом для построения **Системы информационного сопровождения научных исследований нового поколения**. Такая система, с нашей точки зрения, должна состоять из совокупности региональных и тематических информационно-аналитических центров сопровождения научных исследований (ИАЦСНИ), каждый из которых должен иметь три сектора – технологический, научный и технический.

Технологический сектор включает отделения, расположенные в обслуживаемых научных учреждениях, его сотрудники обеспечивают непосредственное информационное обслуживание исследователей данного учреждения (возможен вариант, когда отделение обслуживает несколько научных коллективов, базируясь на отдельной территории). Задачи, которые должен решать этот сектор, подробно перечислены нами в [9], хотя и применительно к научным библиотекам. Поскольку, как мы отмечали, академические библиотеки фактически уже давно выполняют роль информационно-аналитических центров, перечисленные задачи могут быть распространены и на ИАЦСНИ. Среди этих задач – приобретение информационных ресурсов на различных носителях в соответствии с информационными потребностями пользователей; техническая работа по организации подключения пользователей к приобретенным сетевым ресурсам; формирование справочного аппарата по имеющимся ресурсам; поиск и опережающее информирование о появлении в сети или на рынке печатной продукции ресурсов, представляющих интерес для пользователей; предоставление по заказам пользователей копий необходимых им материалов; формирование проблемно-ориентированных баз данных и электронных библиотек по тематике исследований обслуживаемого коллектива; учет и ведение информационных ресурсов, созданных обслуживаемым исследовательским коллективом; проведение библиометрических исследований деятельности обслуживаемого коллектива исследователей.

Научный сектор ставит задачи, разрабатывает технологические и программные решения для Технологического сектора. Он же организует (совместно с Техническим сектором) их внедрение в Технологическом секторе. Наряду с этим Научный сектор проводит исследования информационных потребностей разработчиков, вырабатывает рекомендации по приобретению тех или иных ресурсов, организует работу по анализу «данных обратной связи» с пользователями; решает задачи оптимального управления ресурсами Центра, анализирует (с привлечением специалистов из Технологического сектора) имеющиеся печатные информационные материалы на предмет оцифровки и (или) списания; проводит самостоятельно и участвует в совместных с другими организациями научных работах в соответствии с утвержденными направлениями научных исследований; выдает рекомендации по техническому оснащению отделений Технологического сектора. Научный сектор подготавливает обоснованные заявки в вышестоящие организации на финансирование приобретения информационных ресурсов, технических средств и расходных материалов, необходимых для деятельности Центра.

Технический сектор обеспечивает приобретение, установку и сопровождение технических средств общесистемного и прикладного программного обеспечения, необходимых для работы Научного и Технологического секторов; осуществляет расчеты необходимого количества расходных материалов для работы Технологического и Научного секторов, организует их приобретение и распределение.

Система информационного сопровождения научных исследований, построенная по предлагаемым принципам, с одной стороны, будет опираться на многолетний положительный опыт, накопленный в РАН в области приобретения ресурсов и обслуживания ими ученых, с другой стороны, позволит решить многие проблемы, которые не удавалось решить в академической системе (централизованное обеспечение библиотек техникой и программными средствами, организация сопровождения технических и программных средств, организация единого научного информационного пространства и т.п.).

СПИСОК ЛИТЕРАТУРЫ

1. Каленов Н.Е., Слащева Н.А. Комплектование фондов библиотек: печатные и электронные источники? // Научные и технические библиотеки. – 2013. – № 7. – С. 21-32.
2. Мохначева Ю.В. Российско-белорусское научное сотрудничество: библиометрический анализ текущего состояния и перспектив развития // Информационные ресурсы России. – 2010. – № 5(117). – С. 11-15.
3. Слащева Н.А., Харыбина Т.Н. Библиометрические индикаторы научной деятельности ученых Пущинского научного центра РАН // Информационное обеспечение науки: новые технологии: сборник научных трудов / ред. Н.Е. Каленов. – М.: Научный Мир, 2011. – С. 110-117.
4. Лаврик О. Л. Наукометрический анализ отечественного библиотековедения и библиографоведения // Библиосфера. – 2010. – № 2. – С. 51-59.
5. Трескова П. П. Показатели научной продуктивности и рейтинги академических институтов Уральского отделения РАН // Информационный бюллетень Российской библиотечной ассоциации. – 2011. – № 59. – С. 105-108.
6. Баженов С.Р., Павлов А.И. Основные результаты автоматизации ГПНТБ СО РАН за 2010-2012 гг. // Труды ГПНТБ СО РАН. – 2013. – № 5. – С. 181-200.
7. Трескова П.П., Оганова О.А Этапы формирования и развития библиотечно-информационной системы Уральского отделения РАН // Библиосфера. – 2011. – № 3. – С. 9-16.
8. Опорная телекоммуникационная сеть. – URL: http://www.jscc.ru/backbone.shtml
9. Каленов Н.Е. Задачи и функции библиотек РАН в современных условиях // Информатика и ее применение. – 2012. – Т. 6, № 2. – С. 51-58.

Материал поступил в редакцию 25.02.15.

Сведения об авторе

КАЛЕНОВ Николай Евгеньевич – доктор технических наук, профессор, директор БЕН РАН, зам. председателя Информационно-библиотечного совета РАН
e-mail: nek@benran.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 002 : 001.83 (100)

Л.Э. Миндели, В.А. Маркусова

Библиометрические исследования научного сотрудничества: обзор мировых тенденций*

Обсуждаются отечественные и зарубежные библиометрические исследования, посвященные тенденциям научного сотрудничества на национальном и международном уровнях. Отмечается специфика научного сотрудничества в разных областях знания и влияние на него географической близости. Глобализация науки демонстрируется на результатах представительного обследования по темпам роста расстояния при международном сотрудничестве во всех областях знания, включая общественные и гуманитарные науки.

Ключевые слова: *международное научное сотрудничество, соавторство, статьи, цитируемость, глобализация, расстояние, км, области знаний*

Процессы глобализации науки, наблюдаемые в последнее двадцатилетие, привели к значительному росту количества библиометрических исследований, посвященных анализу тенденций научного сотрудничества. Сам факт научного сотрудничества связан с сущностью научного творчества, которое требует межличностного обмена и достижения консенсуса относительно концепций и идей. Современное состояние информационно-коммуникационных технологий (ИКТ) значительно изменило границы расширения международного научного сотрудничества, которое является важнейшим атрибутом современной науки. Мотивы научного сотрудничества различны и многосторонни, к ним относятся такие важные факторы, как оптимизация финансовых ресурсов; возможность использования дорогостоящего оборудования партнера, его опыта работы; желание работать на передовых направлениях науки. Мотивы эти определяются также социокультурными и личностными особенностями исследователя и его страны обитания.

После окончания Второй мировой войны тенденция научного сотрудничества значительно усилилась. Научное сотрудничество является показателем научных связей между организациями и увеличивает потенциал исследователей для решения сложных проблем, привлекая специалистов с различными навыками и опытом. Одним из наиболее важных катализаторов роста международного сотрудничества был приход «большой науки», в частности, строительство крупных ускорителей, которые, наряду с другими факторами, сыграли свою роль в развитии европейских исследований [1]. Использование дорогостоящего оборудования (например, работы на Большом адронном коллайдере), международный язык науки (английский), решение мультидисциплинарных проблем являются важными факторами, стимулирующими такое сотрудничество. Научное сотрудничество, особенно международное, является предметом многочисленных исследований. Этапы развития научного сотрудничества прослежены в работе [1]. Отмечается, что увеличение количества научных журналов и развитие соавторства, появившиеся почти одновременно с профессионализацией науки, а также распространение все более сложных научных приборов способствовали значительному ускорению темпов сотрудничества [2]. Рост научного сообщества в разных странах также является важным фактором стремления к сотрудничеству. За последние 30 лет было выявлено, что научное сотрудничество, в частности, международная коллаборация, повышает качество научного исследования [3], его импакт и видимость [4, 5].

Исследования показали, что международное научное сотрудничество предоставляет особые возможности малым странам [6]. Профессор В. Гланцел из Католического Университета в Бельгии (Katholieke Universiteit, Leuven) [7] отмечал резкий сдвиг после 1990 г. в моделях сотрудничества СССР/России с Европейским Союзом. Целесообразность использования фракционных (дробных) методов подсчета при оценке научного сотрудничества на основе анализа научной продуктивности (НП) отмечалась в работах, выполненных в Европе и Канаде [8, 9]. Исследования показали, что модели

* Работа выполнена при поддержке РГНФ, грант №14-03-00333

сотрудничества могут существенно различаться в зависимости от научной дисциплины [10].

Одним из показателей роста научного сотрудничества является неуклонное увеличение количества соавторов одной публикации. По данным Отчета Национального научного фонда (ННФ) США «Science and Engineering Indicators» (S&EI–2012)[1], в 1990 г. среднее число соавторов одной статьи США составляло 3,2, а в 2010 г. это число достигло значения 5,6. Количество соавторов значительно различается в зависимости от дисциплины: самый высокий рост соавторства наблюдался в астрономии (с 3,1 до 13,8), за последние 20 лет произошло удвоение числа соавторов в физике (с 4,5 до 10,1). Значительно медленнее этот рост был в общественных науках (от 1,6 соавтора в 1990 г. до 2,1 – в 2010 г.) и в математике (от 1,7 соавтора в 1990 г. до 2,2 – в 2010 г.) (www.nsf.gov).

Существует ряд методов сбора эмпирических данных о международном научном сотрудничестве: участие в международных научно-исследовательских организациях; координация и совместные программы научно-исследовательской деятельности; миграционная мобильность научных кадров; переписка по электронной почте между партнерами по исследованиям, совместное использование ресурсов и оборудования, затраты национальных бюджетов на исследования и разработки (R&D) [1]. Однако наиболее распространенным способом изучения научного сотрудничества на национальном и международном уровне стало исследование соавторства в научных публикациях.

Развитие библиометрических исследований, посвященных научному сотрудничеству, стало возможным благодаря созданию Ю. Гарфилдом (E. Garfield) в 1964 г. Science Citation Index (SCI), золотой юбилей которого отмечался в 2014 г. Наличие в SCI адресов организаций авторов и стран сразу привлекло внимание социологов науки и способствовало использованию этих данных для изучения тенденций соавторства и научного сотрудничества индивидуальных специалистов, организаций и стран.

Общие мировые тенденции научного сотрудничества рассматриваются в Отчете S&EI. По данным этого Отчета за 2014 г., более 60% мирового потока статей в 2012 г. были опубликованы авторами из разных организаций и стран, в то время как в 1997 г. доля таких статей составляла всего половину. Наиболее значительный рост отмечался в международном сотрудничестве: доля соавторов из разных стран выросла между 1997 и 2012 гг. с 16 до 25%. Модели международного научного сотрудничества (МНС) в разных странах различны и отражают культурные, исторические и социальные традиции. Области знаний характеризуются разными темпами роста международного сотрудничества. Самая высокая доля МНС наблюдается в астрономии (56% статей). Науки о земле, компьютерные науки, математика и физика, а также биологические науки имеют долю сотрудничества в пределах от 27 до 34%. В таблице представлены данные о тенденциях соавторства по областям знаний, взятые из вышеупомянутого Отчета.

Обращает на себя внимание значительный рост за последние 15 лет МНС в области общественных наук, хотя в силу традиций этих областей знаний, они несколько отстают от естественных наук.

Доля МНС США составила в 2012 г. 35%, что значительно ниже, чем доля МНС во Франции, Германии и Великобритании. Этот факт связан с очень высокой научной продуктивностью США (26,5% мирового потока), а также с тем, что эта страна имеет более высокую долю статей с отечественными соавторами.

Доля статей, опубликованных при международном научном сотрудничестве по классификации S&EI–2014
(по данным Отчета S&EI – 2014, Fig. 5–22)

Область знания	1997 г.	2012 г.
Астрономия	39,6	56,4
Науки о Земле	20,1	33,7
Компьютерные науки	18,6	30,9
Математика	21,3	30,4
Физика	23,3	28,2
Биология	16,7	27,4
Сельскохозяйственные науки	12,6	23,3
Медицина	11,7	22,2
Технические науки	13,2	21,7
Психология	8,3	20,6
Химия	13,7	20,2
Общественные науки	8,7	19,5
Другие науки	4,5	16,8

[1] Отчет Национального научного фонда США «Science and Engineering Indicators» публикуется каждые два года.

Более высокие темпы роста МНС по сравнению с США имеют экономически сильные страны-члены Европейского Союза (ЕС). Это обстоятельство связано с менее мощной научной базой (как по масштабу персонала, так и по финансовым возможностям), что приводит к необходимости создания совместных научных коллективов с международным участием. Кроме того, рамочные программы ЕС по исследованиям и технологическому развитию и другие программы, направленные на расширение сотрудничества между странами-членами ЕС и с другими странами, в значительной мере способствовали усилению МНС.

Несмотря на фантастический рост научной продуктивности Китая, занимающего с 2007 г. второе место (12,6% мирового потока) в мире, доля его международного сотрудничества составляет не более 15%. Такая же тенденция наблюдается в Японии, доля МНС которого значительно ниже, чем США. Частично это можно объяснить отсутствием в странах Азии формальных рамочных программ для облегчения международного сотрудничества. Другим возможным фактором является недостаточный уровень знания английского языка у китайских и японских ученых, что ограничивает их возможности опубликования результатов исследований в международных научных журналах.

Вопросы международного научного сотрудничества рассматривались во многих работах американского проф. Д. Бивера (D. Bever), который считается в наукометрии «отцом» международной организации COLLNET (www.collnet.org), сообщества специалистов по библиометрии, изучающих тенденции научного сотрудничества как на международном, так и на национальном уровне. В частности, в работе Бивера [11] рассмотрена история научного сотрудничества с 1800 (!) по 2000 г., и проанализированы его положительные и негативные стороны.

Необычный подход в изучении научного сотрудничества изложен в результатах исследования [1], выполненного под руководством доктора Р. Тайссена (R. Tijssen) сотрудниками известного в мире Центра по анализу науки и техники (SWTS, Нидерланды). К проблемам научного сотрудничества с точки зрения глобализации науки они подошли, измеряя физическое расстояние (в км) между странами и организациями.

Анализ был выполнен на аггрегированном уровне стран и областей науки. По мнению авторов, марш глобализации на мировой карте науки отражается в адресах, предоставляемых исследователями в своих публикациях в открытой научной литературе. Были использованы миллионы публикаций, индексированных в БД CWTS, основанной на библиометрической статистике информационных продуктов кампании TP. Для этого эмпирического анализа более 21,4 млн научных публикаций, включенных в Web of Science за 1990–2009 гг., было использовано специальное программное обеспечение, называемое пространственной наукометрией и созданное в 2009 г. Для каждой организации определялись ее географические координаты (ссылка). Специальное внимание было уделено декодировке адресов стран. Всего в массиве публикаций было 39 млн адресов, относящихся к городу или стране. Другие элементы адреса, например, название организации, улицы и почтовый код, были проигнорированы. Для каждого уникального адреса была подсчитана частота его упоминаний в списке адресов массива. Поскольку выполнение геокодирования для всех уникальных адресов оказалось невозможным, то был выбран порог ограничения и оставлено для анализа около 11 000 адресов, наиболее часто упоминаемых.

Используя результаты процедуры геокодирования, авторы рассчитали «географическое расстояние сотрудничества» (GCD) каждой выбранной публикации. Если в статье содержался только один адрес, то GCD публикации определялось как нулевое. В связи с ограничениями процедуры геокодирования координаты некоторых адресов в документе оказались неизвестными. Число таких статей составило 2,3% выбранного массива. Поэтому адреса статей с неизвестными координатами были исключены, GCD было рассчитано на базе оставшихся адресов. На основе соавторства публикаций были выбраны следующие четыре показателя научной глобализации:

- Среднее географическое расстояние сотрудничества (*Medium Global Collaboration Distance* – MGCD): средний GCD всего массива публикаций;
- Процент публикаций со средним и большим расстоянием сотрудничества (*% Medium and Long Distance Collaboration* – MLDC): с GCD более 200 км;
- Процент публикаций с большим расстоянием сотрудничества (*% Long Distance Collaboration* – LDC): с GCD более 1000 км;
- Процент публикаций с очень большим расстоянием сотрудничества (*% Very Long Distance Collaboration* – VLDC): процент публикаций с GCD более 5000 км.

Для анализа сотрудничества по областям знаний массив был классифицирован по следующим четырем направлениям науки: технические науки и технологии (*Engineering Sciences and Technology* – EST); медицина, науки о живой природе и сельскохозяйственные науки (*Medical Sciences, Life Sciences and Agricultural Sciences* – MLA); естественные науки, компьютерные науки и математика (*Natural Sciences, Computer Sciences and Mathematics* – NCM); общественные науки, гуманитарные, литература и искусство (*Social Sciences, Humanities and Arts* – SSHA). Эти области знания были сгруппированы на основе принадлежности научных журналов к предметным категориям по классификации WoS. Географическое расстояние по координатам сотрудничества было рассчитано по программе на сайте www.gpsvisualizer.com/geocoder/.

Было установлено, что расстояние между соавторами в период с 1980 по 2009 гг. увеличилось примерно в 5 раз – с 334 км в 1980 г. до 1553 км в 2009 г. Наблюдались значительные сдвиги в видах сотрудничества – от преимущественно национальной науки к глобализированной. Процесс движения

в сторону глобализации происходил разными темпами в разных регионах и областях знаний. Доля публикаций при сотрудничестве соавторов, находящихся друг от друга на среднем и большом расстоянии, увеличилась в 2009 г. в 4 раза по сравнению с 1980 г. Доля партнерства с очень большим расстоянием выросла в 5 раз, что является свидетельством продолжающегося процесса глобализации науки.

Рост и эволюция мировой науки обусловлены не только социально-экономическими и политическими факторами, но и факторами когнитивной динамики в научных областях, такими как рост биомедицинской науки, нанонаук и информационно-коммуникативных технологий (ИКТ). Были изучены различия в динамике глобализации для четырех широких областей науки. Такие области знания, как естественные науки, компьютерные науки и математика (NCM), были и до сих пор остаются самыми глобализированными из четырех. Это, очевидно, является результатом многолетних традиций, связанных с природой сотрудничества «большой науки» (физика и астрономия), в которой крупные исследовательские установки используются совместно учеными всего мира.

Области медицинских наук (MLA), которые составляли в 1980 г. две трети расстояния (MGCD) в области естественных, компьютерных и математических наук (NCM) к 2009 г. догнали уровень глобализации в этих областях.

Технические науки и технологии (ET), следовали тем же тенденциям роста расстояния международного партнерства, но, начиная с 2003 г., расширение их сети сотрудничества не поспевало за неуклонным темпом мировой глобализации. В последние годы наметился значительный рост партнерства в области социальных наук, гуманитарных наук и искусств (SSHA). Хотя ученые в этих областях знаний традиционно были менее склонны к сотрудничеству с другими исследователями.

Расстояния между партнерами по исследованиям, очевидно, также зависят от географического положения их организации. Те, кто расположен в центре наукоемких стран, регионов или континентов, имеют меньшую потребность в партнерах, находящихся от них очень далеко, чем те, кто работает на географической периферии мировой науки. Географическое расположение страны на земном шаре позволило выявить «периферийные» страны в южном полушарии, характеризующиеся крупнейшими расстояниями по совместной работе. Например, расстояние сотрудничества соавторов из Новой Зеландии составляет более 4000 км. Было также установлено, что некоторые страны в районе тропиков превзошли расстояние в 4000 км. Это, как правило, развивающиеся страны с партнерами в северном или южном полушарии. В работе приведена статистика по пяти странам-лидерам (по расстоянию сотрудничества) в каждой категории расстояния сотрудничества. При выборе стран рассматривались только страны с научной продуктивностью (НП) более 3000 публикаций WoS в 2009 г. Отмечается, что страны, демонстрирующие быстрый рост НП, как правило, имеют очень незначительный или даже отрицательный рост среднего уровня MGCD. Очевидно, что эти страны пытаются добиться быстрого роста НП в основном с помощью публикаций, не связанных с большим расстоянием сотрудничества. Есть основания предполагать, что рост сотрудничества на значительном расстоянии будет наблюдаться позднее, когда эти страны достигнут «глобальной» стадии в развитии национальных научных систем.

Исследование по изучению влияния мультисоавторства и большого количества организаций, вовлеченных в сотрудничество, на его интенсивность [12], было выполнено канадской компанией Science-Metrix, специализирующейся на библиометрических оценках научной деятельности стран, университетов и научных организаций. Для анализа тенденций научного сотрудничества с 2000 по 2009 гг. стран Европейского союза (ЕС), США и стран Африки, была применена методика сложного процесса шкалирования и ряда специальных программ визуализации.

По мнению авторов этого исследования, использование доли МНС не позволит точно установить, почему страны имеют высокую склонность к сотрудничеству, поскольку чем меньше страна, тем больше она тяготеет к сотрудничеству. Избранный в работе метод масштабного шкалирования (рассчитанный по методике компании Science-Metrix) показателей совместной работы, основанный на статистке Web of Science, был использован для 19 стран из группы Большой двадцатки (G20). Интенсивность научного сотрудничества (collaboration intensity) оценивалась по количеству совместных научных работ, и учитывалась общая научная продуктивность каждой из стран. Результат позволил установить, что Индонезия, Германия, Франция, Великобритания, Канада и Италия активно включены в международное научное сотрудничество. Другая тенденция наблюдалась в Турции, странах Азии (Индия, Китай, Япония, Корея), американских странах (Бразилия, США, Аргентина, Мексика), а также в России и Южной Африке. Все эти страны имеют уровень МНС значительно меньше, чем можно было ожидать, учитывая их размер.

Была построена сеть «стремления к сотрудничеству» (collaboration affinity) с использованием библиометрической статистики БД Scopus за 2003–2009 гг. (с применением методики, разработанной компанией Science-Metrix) и выявлена сложная паутина сотрудничества, обусловленного историческими переменными, такими как колониальная история. Например, Франция имеет сильное стремление к сотрудничеству с ее бывшими колониями в Африке (Алжир, Мавритания, Марокко, Ливия и Тунис), та же тенденция наблюдается в сотрудничестве Великобритании, Бельгии и Португалии. Географическая и лингвистическая близость также играют важную роль для стран Латинской Америки. Географическая, культурная и религиозная близость, безусловно, влияет на стремление к сотрудничеству друг с другом мусульманских стран.

Отмечается, что 27 стран ЕС разбросаны по всей сети и не представляют сплоченной группы. Поэтому отдельно было рассмотрено сотрудничество внутри Европейского исследовательского пространства (European research area – ERA). Оказалось, что Люксембург и Бельгия имеют больше совместных статей, чем можно было ожидать (исходя из размера этих стран), что может быть объяснено наличием там европейских институтов. По мнению авторов, многие из стран ERA – это страны, когда-то находившиеся по другую сторону железного занавеса, и ряд из них все еще не сотрудничает внутри ERA настолько, насколько можно было бы ожидать, учитывая их размер (например, Литва, Словения, Польша, Румыния). Однако меньше, чем ожидалось, сотрудничают внутри ERA и такие страны, как Греция, Мальта, Великобритания, Испания, Италия и Германия. И наоборот, Болгария, Словакия и Венгрия активно стремятся к сотрудничеству внутри ERA. Результаты исследования позволили установить влияние географической близости, колониальных и культурных традиций на сотрудничество.

Национальное научное сотрудничество играет важную роль в передаче знаний между различными секторами науки и промышленности. В 1995 г. известным американским проф. Г. Этсковитцем (H. Etzkowitz, Stanford University) и крупным голландским статистиком и специалистом по библиометрии проф. Л. Лейдесдорфом (L. Leidesdorf, University of Amsterdam), было введено понятие «тройной спирали» (Tripple Helix), как одной из характеристик сотрудничества на национальном уровне. Эта концепция является интерпретацией наблюдаемого сдвига от доминирующего партнерства промышленность–правительство в Индустриальном обществе (Industrial Society) к растущему тройному партнерству университеты–промышленность–правительство в Обществе экономики знаний (Knowledge Society) (http://triplehelix.stanford.edu/3helix_concept).

В США, по данным Отчета SEI-2014, наблюдался рост национального научного сотрудничества, доля которого выросла с 36% в 1997 г. до 44% в 2012 г. В этом Отчете представлены сведения о сотрудничестве между академическим сектором (университеты), национальными лабораториями и промышленностью. Отмечается, что именно университеты являются центром научного сотрудничества. Доля статей из университетов, опубликованных со статьями из других секторов научной деятельности, включая международное сотрудничество, составила 53%.

Как отмечается во многих зарубежных публикациях, изучению научного сотрудничества на национальном уровне посвящено недостаточное количество работ. Однако в таком изучении заинтересованы даже чиновники, отвечающие за планирование научной политики и за благосостояние региона. На 14-й международной конференции COLLNET, состоявшейся в г. Тарту (Эстония) в 2013 г., были доложены результаты исследований по национальному научному сотрудничеству, выполненных в Индии, Иране и Турции. Так, рассматривались вопросы национального сотрудничества между элитными научными институтами Индии с визуализацией сети сотрудничества [13]. Изучение влияния централизации научных ресурсов на уровень сотрудничества между университетами и промышленностью в Турции было исследовано на основе соавторства в патентах, относящихся к химической промышленности [14].

В исследовании [15], выполненном в России в 2003–2004 гг. в рамках проекта ИНТАС, были рассмотрены вопросы научного сотрудничества грантодержателей РФФИ. Проанализированный массив публикаций, содержащихся в БД РФФИ, составил 29600 единиц. В этом исследовании внимание было сосредоточено на сотрудничестве между федеральными округами. Была построена карта сотрудничества федеральных округов с использованием системы «Компас-3» (разработанной проф. В.Г. Гитисом). Выполненный в 2012 г. анализ научного сотрудничества отечественных университетов за 2006–2011 гг. по статистике WoS [16] показал, что институты РАН являются центром научного сотрудничества как с сектором высшей школы, так и с другими исследовательскими институтами и национальными федеральными центрами (Курчатовский институт, федеральные ядерные центры и др.). Доля научного сотрудничества федеральных и национальных исследовательских университетов с РАН составила свыше 40%. Несколько неожиданным оказался тот факт, что федеральные университеты между собой практически не сотрудничают, то же самое относится и к национальным исследовательским университетам: этот показатель увеличился с 0,12 % в 2006 г. до 2,0% в 2011 г. Доля сотрудничества федеральных университетов с национальными исследовательскими университетами тоже очень незначительна, хотя и несколько возрастает со временем: с 2,0% в 2006 г. до 2,6% в 2011 г. При этом только в Приволжском федеральном округе расположены шесть НИУ. В то же время доля научного сотрудничества с институтами РАН, например, Сибирского федерального университета и Дальневосточного федерального университета очень высока: в 2011 г. она составила 70,3 и 73,3% соответственно.

В докладе, представленном на 15-й международной конференции COLLNET, состоявшейся в сентябре 2014 г. в г. Ильменау (Германия), отмечалось, что Правительство Китая поставило целью заменить экономику «абсурда» (made in China) на инновационную экономику (innovated in China). В Китае создана правительственная программа «Project 985» по стимулированию научных исследований в ведущих университетах страны, задачей которых является как обучение студентов, так и выполнение фундаментальных научных исследований. В рамках этой программы было исследовано национальное научное сотрудничество среди 39 лидирующих университетов, входящих в «Project 985» [17]. Взаимоотношения научного сотрудничества изучались на основе соавторства статей, которое оценивалось как «частота сотрудничества» (collaboration frequency), т.е. на основе числа совместно опубликованных статей. Временные рамки исследования 2000–2009 гг.

Данные о научной продуктивности (НП) университетов были получены при поиске в самой большой национальной БД Китая (China National Knowledge Infrastructure – CNKI), созданной в 1994 г. и содержащей полнотекстовые копии 6968 китайских научных журналов. Эти научные журналы охватывают все области знания: естественные науки, технические науки, технологии, сельскохозяйственные науки, философию, медицину, гуманитарные и общественные науки. Исследование было выполнено с использованием применяемых в социологии методов и инструментов анализа социальных сетей («Social Network Analysis»). Была построена матрица соавторства и частоты сотрудничества упомянутых 39 университетов. В исследовании были использованы три вида программного обеспечения: Ucinet (инструмент построения матрицы анализа социальной сети), Pajek (визуализация сети) и SPSS (программа статистики, графический анализ, корреляционный и т.д.). В сети научного сотрудничества 39 университетов выявлена 741 пара сотрудничества (совместных статей), что свидетельствует о разветвленной сети сотрудничества. Но иная картина наблюдалась с точки зрения частоты сотрудничества – с максимальным количеством в 1434 совместных статьи и минимальным – в одну. Установлены пары наиболее тесно сотрудничающих университетов.

Если же ввести порог интенсивности сотрудничества (collaboration intensity), то обнаружится, что большинство университетов имеют низкий уровень сотрудничества. При пороге частоты сотрудничества не менее 50 статей две трети университетов выпадут из сети. По мнению авторов, этот факт означает, что отношения сотрудничества многих вузов устанавливаются несколькими авторами. В большинстве случаев такое сотрудничество основано на взаимоотношении руководители–дипломники.

Например, у выпускника университета «А» имеется руководитель из университета «В». Дипломник публикует статью в соавторстве со своим руководителем. В этом случае, статья имеет два адреса организации: университет «А» и университет «В», т.е. два университета имеют отношения научного сотрудничества.

В процессе работы в преддипломной период выпускник, кроме руководителя, может иметь и другого соавтора статьи. Таким образом, этот вид отношений сотрудничества очень хрупок. Как только выпускник получит диплом, взаимоотношение сотрудничества между университетом «А» и университетом «В» будет нарушено. Анализ плотности сотрудничества позволил обнаружить, что большинство университетов из «Project 985» имели хрупкое сотрудничество. Авторы приходят к выводу, что необходимы изменения в научной политике для развития более широкого и интенсивного сотрудничества китайских университетов друг с другом.

Результаты этого исследования демонстрируют общие тенденции научного сотрудничества на национальном уровне, характерные как для Китая, так и для России. Очевидно, что имеется необходимость изменений в научной политике для стимулирования научного сотрудничества внутри сектора высшей школы в обеих странах.

СПИСОК ЛИТЕРАТУРЫ

1. Waltman L., Tijssen R., Eck N. Globalization of science in kilometers // Journal of Informetrics. – 2011. – Vol. 5, № 4. – P. 574–582.
2. Beaver D.B., & Rosen R. Studies in scientific collaboration: Part I. The Professional Origins of Scientific Co-Authorship // Scientometrics. – 1978. – Vol. 1, № 1. – P. 65–84.
3. Presser S. Collaboration and the quality of research // Social Studies of Science. – 1980. – Vol. 10, № 1. – P. 95–101.
4. Glanzel W., & Schubert A. Double effort=double impact? A critical view at international coauthorship in chemistry // Scientometrics. – 2001. – Vol. 50, № 2. – P. 199–214.
5. Glanzel W., Schubert A. & Czerwon H.J. A bibliometric analysis of international scientific cooperation of the European Union (1985–1995) // Scientometrics. – 1999. – Vol. 45, № 2. – P. 185–202.
6. Narin F, Stevens K, Whitlow E. Scientific co-operation in Europe and the citation of multinationally authored papers // Scientometrics. – 1991. – Vol. 21, № 3. – P. 313–23.
7. Braun T., Glanzel W., & Schubert A. Publication and cooperation patterns of the authors of neuroscience journals // Scientometrics. – 2001. – Vol. 51, № 3. – P. 499–510.
8. Glänzel W, Schubert A. Domesticity and internationality in coauthorship, references and citations // Scientometrics. – 2005. – Vol. 65, № 3. – P. 323–342.
9. Archambault E., & Lariviere V. Individual researchers' research productivity: A comparative analysis of counting methods. Paper presented at STI Conference 2010. Leiden, Netherlands.
10. Abramo G., D'Angelo C. A., & Di Costa F. Research collaboration and productivity: is there correlation? // Higher Education. – 2009. – Vol. 57, № 2. – P. 155–171.
11. Beaver D. Reflection on Scientific Collaboration (and its study): Past, Present, Future. Paper presented at the Second Berlin Workshop on Scientometrics and Informetrics. Collaboration in Literature. Berlin, Humbolt University, 2000.
12. Archambault E., Beauchesne O.H., Cote G., Roberge G. Scale-Adjusted Metrics of Scientific Collaboration // Proceedings of the 19th International Conference on Science and Technology Indicators. Leiden, Sept. 3.–Sept. 5, 2014. – URL: http://www.sti2014.cwts.nl/Program/.
13. Nagpaul P.S. Vizualizing the cooperation networks among elite institutions in India // Proceedings of the 9th International conference on We-

bometrics, Informetrics and Scientometrics and 14th COLLNET meeting 2013. Aug. 15–17, 2013. – Estonian Research Council, Tartu. – P. 459–468.

14. Zan B.U., Zan N. The university and industry relation in Turkey: the case of chemistry // Ibid. – P. 34.
15. Markusova V., Minin V., Libkind A., Arapov M., Jansz M., Zitt M. Research in non-metropolitan universities as a new stage of science development in Russia // Scientometrics. – 2004. – Vol. 60, № 3. – P. 365–383.
16. Иванов В.В., Либкинд А.Н., Маркусова В.А. Публикационная активность и научное сотрудничество вузов и РАН // Вестник Российской академии наук. – 2014. – Т. 84, № 1. – С. 30–36.
17. Junping Q., Fangfang W. A Study on the Scientific Research Collaboration Network of "985 Project" Universities in China // Proceedings of the 19th International Conference on Science and Technology Indicators. Leiden, Sept. 3–Sept. 5, 2014. – URL: http://www.sti2014.cwts.nl/Program.

Материал поступил в редакцию 22.01.15

Сведения об авторах

МИНДЕЛИ Леван Элизбарович – член-корр. РАН, доктор экономических наук, профессор, директор ИПРАН РАН, Москва
e-mail: L.Mindeli@issras.ru

МАРКУСОВА Валентина Александровна – доктор педагогических наук, зав. Отделением ВИНИТИ РАН, Москва
e-mail: markusova@viniti.ru

УДК 004.89.032.26

Т.И. Булдакова, Д.А. Миков

Методика анализа информационных рисков с применением нейро-нечёткой сети

Рассмотрена задача анализа информационных рисков в организации. Показано, что процесс анализа рисков должен быть регулярным, предлагается общая методика выполнения анализа рисков, основные её этапы могут быть представлены в виде вложенных процедур (алгоритмов). Начальным этапом методики является анализ циркулирующих в системе потоков данных для реализации предварительных контрмер. Затем выбирается подходящий метод, способный адекватно оценить риски в организации. Предложен метод оценки рисков на основе нейро-нечёткой сети. Приведён пример реализации методики для оценки информационных рисков в среде MATLAB с использованием нейро-нечёткой сети.

Ключевые слова: *защита информации, информационные риски, методы оценки, нейро-нечёткая система, MATLAB*

ВВЕДЕНИЕ

Управление информационной безопасностью играет всё большую роль в деятельности любой организации, использующей современные технологии сбора, хранения и обработки данных. Этот процесс основывается на анализе информационных рисков, под которыми понимают информационные угрозы, их последствия, уязвимость информации и средств её обработки, а также вероятности их возникновения. Необходимо периодически проводить анализ информационных рисков и внедрённых мероприятий по управлению информационной безопасностью, чтобы учесть изменение требований и приоритетов в сфере деятельности организации, либо появление новых угроз и уязвимостей.

В настоящее время существуют различные методики анализа рисков, в которых предлагаются разные способы сопоставления возможных последствий реализации угрозы с вероятностью её реализации и получения соответствующих выводов. Основное отличие представленных методик состоит в выборе шкалы измерения степени риска: количественной или качественной [1-4].

В методиках, использующих количественные методы оценки, риск оценивается через числовое значение, например, в виде размера ожидаемых годовых потерь. При расчёте значений вероятности реализации угрозы, а также уровня возможного ущерба используют, как правило, статистические методы [1, 2]. Однако, если отсутствует достаточное количество статистических данных, то это приводит к снижению адекватности полученных результатов.

Методики, использующие оценку риска на качественном уровне, более распространены [3-5], однако на практике применяют в основном упрощённые качественные методы, позволяющие оценить риск по шкале «высокий», «средний», «низкий». Основной подход – это использование экспертных оценок. Перспективные интеллектуальные технологии для оценки риска пока применяются недостаточно [4, 5].

Цель настоящей статьи – разработка общей методики анализа рисков с учётом реальных условий функционирования системы, обоснование выбора метода оценки риска и повышение адекватности экспертных оценок для настройки используемой нейро-нечёткой сети.

ОСОБЕННОСТИ ПРОЦЕССА АНАЛИЗА ИНФОРМАЦИОННЫХ РИСКОВ

Существуют разные определения информационного риска, не меняющие его сути. Так, в соответствии с [5], риск информационной безопасности R – это комплексная величина, определяемая как функция (или функционал) ряда факторов, например:

$$R = f\left(X_1, X_2, X_3\right),$$

где X_1 – угрозы информационной безопасности; X_2 – потенциально возможный ущерб; X_3 – уязвимости информационной системы.

К основным сложностям, возникающим при проведении анализа риска, относятся:

1) неполнота информации о составляющих риска и их неоднозначные свойства;

2) сложность создания модели информационной системы и оценки её уязвимости;

3) длительность процесса оценки и быстрая потеря актуальности её результатов;

4) сложность агрегации данных из различных источников, в том числе статистической информации и экспертных оценок;

5) необходимость привлечения нескольких специалистов по анализу рисков для повышения адекватности оценок.

В общем случае задача заключается в выборе из множества Y методов выявления и анализа риска информационной безопасности такого метода y^*, который обеспечивал бы максимальную вероятность выявления и адекватной оценки риска с учётом адаптивности к качественным данным о факторах X_1, X_2 и X_3.

Необходимо найти

$$y^* \in Y \Leftrightarrow p_1^* = \max p_1\left(X, p_2(y)\right), \qquad (1)$$

где X – множество факторов информационного риска; p_1 – вероятность выявления и адекватной оценки риска; p_2 – показатель адаптивности метода к качественным данным.

Однако решение поставленной задачи связано с рядом проблем:

1) оценить показатель p_1, для чего нужно знать X;

2) сформировать X с учётом тех факторов риска, которые могут проявиться в реальных условиях функционирования системы;

3) обеспечить достаточное значение показателя p_2.

Поэтому целесообразно поставить задачу шире: требуется общая методика проведения анализа рисков, учитывающая указанные ограничения и сложности, при этом метод оценки рисков должен отвечать требованию (1).

Например, для определения множества X с учётом угроз, ущерба и уязвимостей, проявляющихся в реальных условиях функционирования информационной системы (ИС), должен быть проведён анализ циркулирующих в ней потоков данных. Поэтому этот этап должен быть включён в общую методику анализа рисков, однако, в существующих методиках он отсутствует [1].

Кроме того, необходимо рассмотреть и проанализировать множество Y для оценки эффективности методов выявления и анализа риска информационной безопасности.

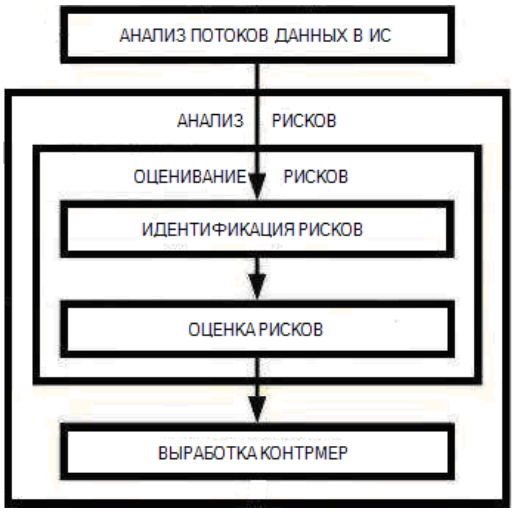
ОСНОВНЫЕ ЭТАПЫ РЕАЛИЗАЦИИ МЕТОДИКИ

Основные этапы реализации методики анализа рисков могут быть представлены в виде вложенных процедур (алгоритмов), представленных на рисунке.

Анализ циркулирующих в ИС потоков данных может быть эффективно выполнен при использовании современных структурных методов, например, с помощью методологии IDEF0 путём разработки и анализа функциональной модели системы [6-7]. В этом случае процесс определения множества X состоит из следующих этапов:

1) производится структурный анализ информационной системы;

2) разрабатывается функциональная модель информационной системы;



Основные этапы методики анализа информационных рисков

3) анализируются циркулирующие в информационной системе потоки данных;

4) определяется полный перечень уязвимостей (используются как типовые, так и специфичные для конкретной системы);

5) на основании функциональной модели и перечня уязвимостей аналитически разрабатывается перечень предварительных контрмер;

6) функциональная модель дорабатывается и конвертируется в ИС с учётом изменений, необходимых для реализации предварительных контрмер.

Пример реализации этапа анализа циркулирующих потоков данных приведён в работах [8-10] применительно к системе электронного здравоохранения [11].

Процесс оценивания рисков состоит из этапов идентификации и оценки рисков.

Идентификация рисков – это процесс нахождения, составления перечня и описания элементов риска. Цель идентификации риска – определить, что могло бы произойти при нанесении возможного ущерба, и получить представление о том, где, как и почему мог иметь место ущерб.

Оценка рисков – это процесс присвоения значений вероятности и последствий риска. Для установления количественной или качественной оценки используется шкала с числовыми либо лингвистическими значениями, соответственно, как последствий, так и вероятности, с применением данных из различных источников [1, 5]. Обычно в дополнение к количественной оценке используют качественную шкалу (например, «низкий», «средний», «высокий»), атрибуты которой будут соответствовать определённым отрезкам на количественной шкале.

Если информация, полученная в результате оценки, достаточна для определения действий, необходимых для снижения информационных рисков до приемлемого уровня, то можно переходить к этапу выработки итоговых контрмер.

В противном случае следует вернуться к первому этапу – анализу потоков данных в ИС, чтобы скорректировать, уточнить и дополнить исходные данные.

На этапе выработки контрмер выбираются варианты обработки риска, и определяется остаточный риск. Существует четыре типовых варианта обработки риска:

1) минимизация риска (выполнение действий для минимизации вероятности и/или негативных последствий, связанных с риском);

2) принятие риска (готовность организации понести ущерб от конкретного риска в случае, если его уровень считается допустимым);

3) уклонение от риска (отказ от вовлечения в рискованную ситуацию или действие, предупреждающее её возникновение);

4) передача риска (перенесение ответственности за риск на третьи лица).

Таким образом, процесс анализа рисков информационной безопасности в организации необходимо начинать с анализа потоков данных в ИС, который состоит из структурного анализа информационной системы и разработки функциональной модели. Затем нужно выбрать подходящий метод, способный адекватно оценить риски в организации.

ВЫБОР МЕТОДА ОЦЕНКИ ИНФОРМАЦИОННЫХ РИСКОВ

Чтобы оценить эффективность выявления и анализа рисков, необходимо рассмотреть множество методов Y, которые можно разделить на три основные группы:

1) статистические методы;

2) методы экспертных оценок;

3) методы моделирования.

Статистические методы предполагают анализ уже накопленных данных о реально случавшихся инцидентах, связанных с нарушением информационной безопасности. На основе результатов такого анализа строятся предположения о вероятности проведения атак и уровнях ущерба от них в аналогичных информационных системах. Эти методы достаточно распространены и просты в применении. Однако их использование не может считаться серьёзным решением из-за неполноты и, зачастую, неточности накопленных статистических сведений, а также в силу их неспособности учитывать скрытые уязвимости, с которыми не был связан ни один инцидент информационной безопасности, но которые могут стать причиной инцидентов в будущем. Поэтому показатель вероятности выявления и правильной оценки риска p_1 статистическими методами не превышает 0,1-0,3 в зависимости от полноты статистической информации и наличия внутренних данных о самой организации, что позволяет несколько повысить p_1 [1, 12].

При использовании методов экспертной оценки анализируются результаты работы группы экспертов, компетентных в области информационной безопасности, которые на основе имеющегося у них опыта определяют количественные или качественные уровни информационных рисков. Методы экспертной оценки являются наиболее распространёнными, но при этом они характеризуются достаточной субъективностью, непрозрачностью и невозможностью проверить экспертное мнение. Поэтому неизбежно возникает вопрос об адекватности оценок и решений, предлагаемых экспертами. Вероятность p_1 выявления и правильной оценки риска методами экспертных оценок выше, чем в случае со статистическими методами, и составляет 0,4-0,6, но всё равно недостаточна, поэтому эту группу методов рекомендуется применять только на этапе оценки входных данных – факторов риска (угроз, потенциально возможного ущерба и уязвимостей). В этом случае для выявления степени согласованности экспертов (и повышения показателя p_1) целесообразно использовать коэффициент конкордации [13, 14].

Методы моделирования основаны на построении, изучении и анализе математических моделей, описывающих функционирование информационной системы. Применение подобных моделей позволяет проанализировать и оптимизировать процессы сбора, хранения и обработки информации, а также выбрать технологии защиты данных [13, 15, 16]. Математическая модель системы, отражая физическую суть процессов её функционирования, позволяет адекватно оценить различные характеристики ИС. Кроме того, в последнее время для построения моделей сложных процессов и систем эффективно применяют интеллектуальные технологии [4, 17, 18]. Показатель p_1 у методов моделирования выше, чем у остальных, и составляет 0,8-0,9, поэтому целесообразно именно их использовать для оценки рисков. Анализ методов моделирования и обоснование эффективности применения нейро-нечётких моделей для оценки рисков выполнены в работах [13, 18, 19]. Более того, в работе [4] экспериментально доказано, что большой адаптивностью к качественным данным (показатель p_2) обладают именно методы «мягкого» моделирования, среди которых наибольшую адаптивность продемонстрировали гибридные модели на основе нечёткой логики.

Поэтому можно сделать вывод, что наибольшими показателями эффективности обладают нейро-нечёткие сети (ННС), которые способны выявлять и адекватно оценивать риск информационной безопасности за счёт нейросетевого компонента, а также за счёт использования нечёткой логики они адаптивны к нечисловым данным.

РАЗРАБОТКА НЕЙРО-НЕЧЁТКОЙ СЕТИ В СРЕДЕ MATLAB

Выбор архитектуры сети и принципы её обучения подробно рассмотрены в работе [19]. Для практической работы с нечёткой логикой и задания функций принадлежности существуют специализированные программы, облегчающие этот процесс. Поэтому для реализации нейро-нечёткой сети воспользуемся редактором FIS программного комплекса MATLAB, который обладает графическим интерфейсом и позволяет вызывать все другие редакторы и программы просмотра систем нечёткого вывода [20].

Для создаваемой нечёткой модели выбраны следующие параметры:

1) три входные (угроза, ущерб, уязвимость) и одна выходная (риск) переменные;

2) тип системы нечёткого вывода – Сугено (для создания ННС);

3) And method (Метод логической конъюнкции) – prod (метод алгебраического произведения);

4) Or method (Метод логической дизъюнкции) – probor (метод алгебраической суммы);

5) Implication (Метод вывода заключения) – min (метод минимального значения);

6) Aggregation (Метод агрегирования) – max (метод максимального значения);

7) Defuzzification (Метод дефаззификации) – wtaver (метод взвешенного среднего).

Для трёх входных переменных (угроза, ущерб, уязвимость) выбрано пять нечётких классов (очень низкий, низкий, средний, высокий, очень высокий) и трапецеидальная функция принадлежности.

Для выходной переменной (риск) выбрано девять нечётких классов (пренебрежимо низкий, очень низкий, низкий, ниже среднего, умеренный, выше среднего, высокий, очень высокий, критический), которые в нечёткой системе типа Сугено принимают фиксированные значения на отрезке [0, 1].

Итак, система нечёткого вывода содержит три входные переменные с пятью термами, 125 правил нечётких продукций и одну выходную переменную с девятью термами.

По оценкам входных переменных определяется риск, отражающий фактическое состояние защищённости исследуемой системы на данный момент времени. Для настройки сети могут быть использованы экспертные оценки, для выявления их согласованности целесообразно применить коэффициент конкордации.

В качестве иллюстрации можно взять одну из характеристик, полученных в результате исследования IDEF0-модели информационной системы виртуального центра охраны здоровья (ВЦОЗ) [11]. Например, существует проблема отслеживания действий каждого клиента в произвольный момент времени, чтобы выявить потенциального нарушителя. Необходимо предположить, что на основе предварительного обследования получены некоторые оценки вероятности реализации угрозы (нарушение клиентом политики безопасности), величины потенциально возможного ущерба (урон, который понесёт ВЦОЗ в результате совершённого нарушения) и степени уязвимости (отсутствие протокола взаимодействия «клиент-сервер», не позволяющее своевременно отследить и выявить нарушителя). Например, угроза – 0,68, ущерб – 0,74, уязвимость – 0,72. Тогда риск равен 0,745, что соответствует значению «высокий» по шкале уровней риска.

Построенная нейро-нечёткая сеть имеет гибкие настройки, удобна и проста в применении, а также точно и наглядно отображает зависимость уровня информационного риска от значений угроз информационной безопасности, потенциально возможного ущерба и уязвимостей информационной системы.

Программа просмотра поверхности системы нечёткого вывода позволяет просматривать поверхность системы нечёткого вывода и визуализировать графики зависимости выходных переменных от отдельных входных переменных.

Алгоритм оценки информационных рисков на основе применения разработанной сети состоит из следующих этапов:

1. Проведение экспертного опроса для получения оценок мощности угрозы (a_1), величины ущерба (a_2) и степени уязвимости (a_3) в интервале [0, 10].

2. Обеспечение адекватности экспертных оценок через вычисление коэффициента конкордации:

$$W = \frac{12S}{n^2 \times \left(m^3 - m\right)}.$$

Здесь W – коэффициент конкордации, S – сумма квадратов отклонений сумм оценок (ответов, данных всеми экспертами на каждый вопрос) от среднего арифметического сумм оценок, n – число экспертов (число ответов на один вопрос), m – число вопросов. Коэффициент конкордации W лежит в границах [0, 1]. Чем ближе значение коэффициента к единице, тем выше уровень согласованности мнений экспертов. Обычно минимально допустимое значение коэффициента конкордации составляет 0,4. Поэтому при согласованном результате $W \geq 0{,}4$ [9-11].

3. По оставшимся оценкам производится вычисление входных переменных ННС – максимальных значений вероятности реализации угрозы информационной безопасности (x_1), нанесения наивысшего возможного ущерба (x_2) и использования уязвимости информационной системы (x_3). Так как под переменными x_1, x_2, x_3 понимаются вероятности, то их значения должны быть в интервале [0, 1]:

$$\begin{cases} 0 \leq x_1 \leq 1, \\ 0 \leq x_2 \leq 1, \\ 0 \leq x_3 \leq 1. \end{cases}$$

Итоговая система уравнений

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \leq b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \leq b_2, \\ \quad\quad\quad \cdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 \leq b_n, \end{cases}$$

$$a_{ij} \in [0,10]; b_i \in [0,30]; i \in 1,2,\ldots,n; \; j \in 1,2,3,$$

решается симплекс-методом. Здесь a_{i1} – экспертные оценки мощности угрозы; a_{i2} – экспертные оценки величины ущерба; a_{i3} – экспертные оценки степени уязвимости; b_i – оценки риска; n – число экспертов.

4. Подача полученных значений переменных x_1, x_2, x_3 на вход разработанной ННС.

5. Получение значения уровня риска информационной безопасности, сопоставление с качественной шкалой, анализ результатов и выработка контрмер на основе проведённого анализа.

Таким образом, алгоритм оценки информационных рисков включает проведение экспертного опроса для получения предварительных оценок, обеспечение адекватности экспертных оценок через вычисление коэффициента конкордации и отсеивание крайних значений, а также вычисление входных переменных нейро-нечёткой сети на основе оставшихся экспертных оценок, подачу полученных значений на вход нейро-нечёткой сети и выработку контрмер на основе анализа полученной выходной переменной.

ЗАКЛЮЧЕНИЕ

Методика оценки и анализа рисков информационной безопасности на основе нейро-нечёткого моделирования позволяет учитывать качество входной информации и надёжность (степень доверия) источников информации. Методика обладает широкими возможностями, позволяющими адаптировать её к разнообразным профилям прикладных систем и встраивать в состав собственных разработок систем управления рисками.

СПИСОК ЛИТЕРАТУРЫ

1. Астахов А.М. Искусство управления информационными рисками. – М.: ДМК Пресс, 2010. – 312 с.
2. Велигура А.Н. О выборе методики оценки рисков информационной безопасности // Information Security/ Информационная безопасность. – 2008. – № 4. – С. 16-17.
3. Rot A. IT Risk Assessment: Quantitative and Qualitative Approach // Proceedings of the World Congress on Engineering and Computer Science. – 2008. – P. 1073-1078.
4. Lee Ming-Chang. Information Security Risk Analysis Methods and Research Trends: AHP and Fuzzy Comprehensive Method // International Journal of Computer Science & Information Technology (IJCSIT). – 2014. – Vol 6, № 1. – P. 29-45.
5. Атаманов А.Н. Модуль нечеткого вывода на основе нейронных сетей для динамического итеративного анализа рисков информационной безопасности // Безопасность информационных технологий. – 2011. – № 1. – С. 7.
6. Калянов Г.Н. CASE-технологии. Консалтинг при автоматизации бизнес-процессов. – М.: Горячая линия – Телеком, 2002. – 320 с.
7. Булдакова Т.И. Проектирование информационных систем управления. – Саратов: Поволжская академия государственной службы им. П.А. Столыпина, 2007. – 160 с.
8. Миков Д.А. Построение IDEF0-модели виртуального центра охраны здоровья // Молодёжный научно-технический вестник. – 2013. – № 09. – С. 37.
9. Булдакова Т.И., Миков Д.А. Анализ информационных процессов виртуального центра охраны здоровья // Научно-техническая информация. Сер. 2. – 2014. – № 2. – С. 10-20.
10. Булдакова Т.И., Суятинов С.И., Миков Д.А. Анализ информационных рисков виртуальных инфраструктур здравоохранения // Информационное общество. – 2013. – № 4. – С. 6
11. Анищенко В.С., Булдакова Т.И., Довгалевский П.Я. и др. Концептуальная модель виртуального центра охраны здоровья населения // Информационные технологии. –2009. – № 12. – С. 59-64.
12. Лагутин М.Б. Наглядная математическая статистика: учеб. пособие. – М.: БИНОМ. Лаборатория знаний, 2007. – 472 с.
13. Булдакова Т.И., Миков Д.А. Метод повышения адекватности оценок информационных рисков // Инженерный журнал: наука и инновации. – 2012. – № 3 (3). – С. 36.
14. Миков Д.А. Управление информационными рисками с использованием экспертного опроса. – Германия, Саарбрюккен: LAP LAMBERT Academic Publishing, 2013. – 83 с.
15. Шаньгин В. Ф. Информационная безопасность компьютерных систем и сетей. –М.: ИД «ФОРУМ»: ИНФРА-М, 2008. – 416 с.
16. Булдакова Т.И., Джалолов А.Ш. Анализ информационных процессов и выбор технологий обработки и защиты данных в ситуационных центрах // Научно-техническая информация. Сер. 1. – 2012. – № 6. – С. 16-22.
17. Булдакова Т.И. Нейросетевая защита ресурсов автоматизированных систем от несанкционированного доступа // Наука и образование: электронное научно-техническое издание. – 2013. – № 05. – С. 269-278.
18. Булдакова Т.И., Джалолов А.Ш. Особенности разработки интеллектуальной системы защиты информации в ситуационном центре // Научно-техническая информация. Сер. 2. – 2014. – № 4. – С. 1-8.
19. Булдакова Т.И., Миков Д.А. Оценка информационных рисков в автоматизированных системах с помощью нейро-нечеткой модели // Наука и образование: электронное научно-техническое издание. – 2013. – № 11. – С. 295-310.
20. Леоненков А.В. Нечёткое моделирование в среде MATLAB и fuzzyTECH. – СПб.: БХВ-Петербург, 2005. – 736 с.

Материал поступил в редакцию 18.08.14.

Сведения об авторах

БУЛДАКОВА Татьяна Ивановна – доктор технических наук, профессор кафедры «Информационная безопасность» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: buldakova@bmstu.ru

МИКОВ Дмитрий Александрович – аспирант кафедры «Информационная безопасность» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: MikovDA@yandex.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК [811.161.1 : 811.581]’322.2

Ю. Тао, В.П. Захаров

Разработка и использование параллельного корпуса русского и китайского языков*

Описывается создание самого большого в Китае параллельного русско-китайского корпуса. Приводятся сведения об общей структуре корпуса, дается характеристика отобранных текстов, показан процесс создания корпуса, а также автоматическая генерация словника терминов. Рассмотрены возможности поиска словоизменительных парадигм. Представлены примеры лингвистического анализа на основе корпуса.

Ключевые слова: *параллельный корпус русского и китайского языков, корпусный менеджер, поиск, генерация словника терминов, переводоведение, контрастивная лингвистика, преподавание*

ВВЕДЕНИЕ

В последние* годы создание корпусов и корпусно-ориентированные исследования стали неотъемлемой частью деятельности лингвистов. Корпусная методология становится частью лингвистической науки, и все лингвисты, работающие в самых разных областях, как правило, проводят свои исследования на базе корпусов.

Одно из направлений корпусной лингвистики – создание и использование параллельных корпусов, которые применяются для решения разнообразных задач, таких как создание и настройка систем машинного перевода, сравнительное изучение языков, развитие теории переводоведения, обучение языкам [1-3]. Корпусы и конкордансы к ним предоставляют лингвистам, переводчикам, переводоведам и студентам бесценный и ранее недоступный лингвистический материал, характеризующийся большим объемом, разнообразием стилей и жанров, с возможностью быстрого нахождения примеров на анализируемые слова и конструкции.

В настоящей работе мы представляем процесс разработки параллельного корпуса русского и китайского языков и возможные направления исследований на его основе. Это первый параллельный корпус русского и китайского языков в Китае. Разработка проекта корпуса, формирование коллекции текстов, проведение предварительной обработки, создание поисковой системы и паспорта корпуса – все это составляло цепь непростых задач. Как мы преодолевали эти трудности и создавали корпус? Чем он отличается от других корпусов? Каковы возможности его использования?

КОРПУСНАЯ ЛИНГВИСТИКА И ПАРАЛЛЕЛЬНЫЕ КОРПУСЫ

Еще в 1897-98гг. немецким лингвистом Кедингом был составлен первый корпус текстов в бумажном виде для сравнения частоты распределения букв в словах и выявления их сочетаемости [4].Однако технологию и результаты исследования нельзя было назвать перспективными в силу того, что вручную проанализировать такое большое собрание текстов – задача практически непосильная. Появление компьютеров позволило решить эту проблему. Разработки программного обеспечения для работы с корпусами текстов привели к созданию специальных программ-*конкордансеров*, которые впоследствии стали называть *корпусными менеджерами* [5, с. 10-11]. Название *конкордансер* прямо указывает на основное назначение этих программ, а именно, выдачу конкордансов – определенным образом упорядоченных списков искомых слов с некоторым контекстным окружением (старое, близкое название из информационного поиска – KWIC, Key Word in Context).

Существует множество определений термина *корпус*. Все они так или иначе фиксируют основные компоненты этого понятия. Корпус должен быть

* Исследование поддержано грантом Бюро национального фонда социальных и гуманитарных наук Китайской Народной Республики № 13BYY026 «Исследование перевода тематических текстов на основе параллельного корпуса русского и китайского языков».

электронным, репрезентативным, размеченным и включать тексты и фрагменты текстов, отобранные по определённому принципу в соответствии с четкими языковыми критериями, определяемыми решаемой задачей (см., например, [6, с. 3; 7, с. 21; 8, с. 5]).

Практика разработки и применения электронных корпусов текстов показала, что невозможно создать универсальный корпус. Задачи и цели любого исследования, которое предполагается проводить с помощью корпусов, определяют тип корпуса, правила отбора текстов, способ и степень их обработки. В области корпусной лингвистики уже создано огромное число корпусов, предназначенных для различных типов исследований, и задача их классификации требует выделения разнообразных характеристик – оснований для классификации [8, с. 16-19]. При этом практически во всех классификациях присутствует деление корпусов на одноязычные и многоязычные. Двуязычные и многоязычные корпусы, в свою очередь, можно разделить на два основных типа:

1) параллельные, или переводные, корпусы (parallel, translation corpora), представляющие множество текстов-оригиналов, написанных на каком-либо исходном языке, и переводов этих исходных текстов на один или несколько других языков; правда, существуют корпусы, в которых все языки признаются равнозначными, например, корпусы, созданные на основе официальных документов ООН или Европейского Сообщества;

2) псевдопараллельные, или сопоставимые, корпусы (comparable corpora), собранные по сходным критериям, объединяющие тексты из одной и той же тематической области, написанные на двух или нескольких языках, но тексты оригинальные, не являющиеся переводами. Сопоставимые корпусы, написанные на одном языке, могут служить «фоном» для параллельного корпуса. Так, в нашем случае, исследуя перевод с русского языка на китайский, мы обращаемся к сопоставимому корпусу на китайском, чтобы выявить особенности китайского языка в переводных и непереводных текстах.

Корпусы обоих типов используются для разработки эффективных методов перевода, а также для сравнительных исследований языков. Они позволяют идентифицировать те или иные приемы перевода, оценить их эффективность, проанализировать лексику и грамматику текста перевода в сопоставлении с оригинальным текстом, сравнить и оценить различные стратегии перевода, найти на основе списков контекстов соответствия тем или иным стилистическим явлениям и выделить способы их передачи при переводе. Параллельные корпусы являются своего рода сборниками стратегий и эквивалентов перевода, которыми руководствовались и которые придумывали переводчики. Они обеспечивают нас информацией, которую двуязычные словари обычно не содержат. Они предлагают эквиваленты не только на уровне слова, но и на уровне конструкций и словосочетаний, а также переводы безэквивалентной лексики. Параллельные и псевдопараллельные корпусы используются также при создании систем машинного перевода и как ресурс для автоматического извлечения терминов и терминологических словосочетаний данной предметной области для нескольких языков [9].

Параллельные корпусы открывают возможности для компаративистских исследований, дают новую информацию по сравнению с исследованиями на базе одноязычных корпусов [1, с. 12], расширяют наши знания о языках, их универсальных особенностях наряду с типологическими и культурными различиями.

В Китае история развития корпусной лингвистики условно может быть поделена на три этапа: до 1980 г., с 1980-го до середины 1990-х гг. и после середины 1990-х гг. [10]. В первый период методология корпусной лингвистики применялась к корпусам в виде печатных текстов с их ручной обработкой. На втором этапе формировались электронные корпусы (объемом порядка 1 млн иероглифов), на основе которых создавались частотные словари и учебные пособия. Важно отметить, что на этом этапе был создан национальный стандарт сегментации текста на слова (The Segmentation Criterion for Modern Chinese Used for Information Processing, GB-13715, October of 1990) [10]. С середины 1990-х компьютеры стали использоваться весьма широко и начали создаваться корпусы большого объема и разных типов. Появились также параллельные корпусы для китайского языка, но вторым языком почти всегда выступал английский. В последние годы было создано несколько параллельных англо-китайских корпусов [10, 11], китайско-английских [12] и японско-китайских корпусов [13], на основе которых проводился анализ универсалий [14, с. 51-52] и норм перевода.

Параллельных русско-китайских или китайско-русских корпусов до сих пор практически не существует. В настоящее время в Китае создается еще один параллельный корпус русского и китайского языков, но только на основе текстов по военной тематике [15]. Поэтому создание большого тематического (в широкой гуманитарной области) параллельного русско-китайского корпуса имеет большое значение.

ЦЕЛИ И ЗАДАЧИ СОЗДАНИЯ КОРПУСА

Цель разработки корпуса – создать программно-лингвистическую платформу для исследований в области перевода с русского языка на китайский и для обучения русскому языку.

Теоретическое значение разработки корпуса состоит в том, что он дает возможность на большом объеме текстов провести сравнительный анализ лексических и грамматических средств выражения в двух языках. Между китайским и русским языками существует большая разница, в первую очередь, в морфологии. В отличие от русского языка с богатой морфологией (словоизменение с категориями рода, числа и падежа у имен существительных, времени, вида и наклонения – у глаголов и т.д., словообразование различного типа) в китайском языке морфология фактически отсутствует [16, 17].

Также большие отличия наблюдаются и в синтаксисе, особенно в способах построения сложного предложения. Особенность китайского языка – паратаксис, в китайском языке совсем нет (мало) формальных средств связи простых предложений в составе сложного. Современные китайские учёные в

связи с этим даже утверждают, что в китайском языке «семантика занимает центральное место» [18, с. 1] и «семантика определяет грамматику» [19, с. 173].

Будучи уверенными, что китайский язык в переводах с русского обнаруживает как типологию соответствий, так и своеобразие своей природы, мы хотим проверить, превалирует ли семантический компонент в переводных текстах, и выявить на большом корпусном материале количественные параметры переводческой практики.

Корпусный подход породил новую научно-исследовательскую парадигму в китайском переводоведении. Он внес существенный вклад в изучение универсалий перевода, впервые описанных М. Бейкер [20, 21] и указывающих на характерные особенности переводных текстов по сравнению с оригинальными, написанными на том же языке (их еще называют «спонтанные»). Утверждается, что язык перевода может обладать какими-то специфическими чертами именно потому, что это перевод. Это особенно важно исследовать, учитывая специфику китайского языка, когда переводчикам часто приходится искать способы преодоления различий в средствах выражения между китайским языком и языком оригинала.

Согласно [22], главные универсалии перевода представлены упрощением, экспликацией, стандартизацией (simplification, explicitation, standardization). Поясним коротко эти понятия. *Упрощение* – стратегия перевода, в результате которой переводные тексты в среднем выглядят проще с точки зрения лексики и синтаксиса, чем оригинальные. При *экспликации* в переводном тексте появляются дополнительные черты, свидетельствующие о попытках переводчика пояснить оригинальный текст, сделать его более понятным. *Стандартизация* (или *нормализация*) – это стратегия, ведущая к тому, что в переводных текстах употребляется меньше «нестандартных» слов и конструкций, чем в оригинальных (меньше синонимов, регулярная нормализация окказиональной лексики и т. п.).

И наличие двух корпусов – корпуса переводов на китайский и сопоставимого корпуса на том же языке – дает нам возможность выявить наличие или отсутствие расхождений в переводных текстах и в «спонтанных», относящихся к тому же жанру и к той же тематике, и попытаться ответить на вопрос, какие же универсалии перевода наблюдаются в области русско-китайского перевода в гуманитарной сфере.

Конкретные задачи разработки корпуса следующие:

1) исследование переводов;

2) преподавание русского языка: внедрение новой формы обучения, базирующейся на больших массивах текстов, создание учебных материалов нового типа.

ПОСТАНОВКА ЗАДАЧИ

Наш корпус характеризуется следующими параметрами:

- объект исследования – русский и китайский языки;
- корпус содержит тексты гуманитарных и общественных наук из следующих областей: (1) политика и международные отношения; (2) лингвистика; (3) литературоведение; (4) переводоведение;
- большой объем корпуса;
- предусмотрено пополнение корпуса новыми текстами;
- корпус будет представлен в открытом доступе;
- корпус содержит в своем составе сопоставимый корпус по той же тематике.

Это специальный и «гомогенный» корпус [23, с. 78], содержащий тексты только гуманитарной области. С точки зрения практики перевода в Китае в прошлом веке большинство переводных произведений составляли художественные и технические тексты, в этом же веке востребован и быстро развивается перевод научных гуманитарных текстов. И поэтому важно анализировать особенности перевода в этой области.

Научный стиль в целом и в отдельных предметных областях имеет ряд особенностей, что даёт возможность говорить о его специфике. Научный стиль, как известно, характеризуется логической последовательностью изложения, упорядоченной системой связей между частями высказывания, стремлением авторов к точности, сжатости, однозначности при сохранении насыщенности содержания. Особые лексические единицы, термины, точно и однозначно называют специальные понятия научной сферы и раскрывают их содержание. Стремление к информационной насыщенности обусловливает отбор наиболее емких и компактных синтаксических конструкций.

Если текст на исходном языке обладает какими-то особенностями, то выражает ли это эксплицитно язык перевода? Проявляется ли нормализация при переводе терминов? Какие методы применяются в процессе перевода сложных, но однотипных синтаксических конструкций, характерных для научного стиля? Мы отвечаем на эти вопросы, применяя методологию корпусного переводоведения.

В процессе работы над корпусом предстояло выполнить следующее:

а) провести отбор и начальный ввод текстов;

б) выполнить их метаописание;

в) создать или адаптировать модуль поисковой системы (корпусный менеджер);

г) загрузить тексты в корпус;

д) провести статистический анализ корпусных данных;

е) разработать пользовательскую документацию;

ж) провести экспериментальную эксплуатацию.

з) проанализировать полученные результаты и разработать программу развития корпуса.

СОЗДАНИЕ КОРПУСА

Структурирование и объем корпуса

Корпус состоит из двух частей:

1) параллельный корпус текстов на русском языке и их переводов на китайский язык (parallel corpus – PC);

2) сопоставимый корпус (непереводные тексты) (comparable corpus – CC), состоящий из текстов на китайском языке, тема и содержание которых подобны текстам в параллельном корпусе (рис. 1).

На начальном этапе суммарный объем параллельного корпуса – более 3 млн слов, сопоставимого – около 1 млн слов. Оба подкорпуса будут открытыми и динамическими. Корпус будет пополняться каждые два года, при этом будут развиваться техника разметки и программное обеспечение.

Параллельный корпус (РС) включает в себя 10 монографий на русском языке и их переводы на китайский язык из области политики, международных отношений, лингвистики и литературоведения[1] (*Приложение 1*). Эти монографии были переведены в последние тридцать лет и являются текстами современного русского языка.

Сопоставимый корпус (СС) включает в себя 10 монографий на китайском языке из тех же предметных областей (*Приложение 2*). Совпадение жанра и тематики текстов параллельного и сопоставимого корпусов обеспечивает возможность изучения особенностей переводного языка в сравнении с оригинальным китайским языком. Переводной китайский язык сопоставляется как с оригинальным русским языком в параллельном корпусе, так и с оригинальным китайским языком в сопоставимом корпусе. Оригинал и перевод каждой монографии параллельного корпуса создают свой подкорпус, а каждая монография сопоставимого корпуса образует соответствующий подкорпус. Таким образом, мы имеем подкорпусы РС1, РС2, РС3...РС10 и СС1, СС2, СС3...СС10. И сопоставление, и количественный анализ идет не только между корпусами, но и между парами подкорпусов.

Для обеспечения достоверности исследования мы постарались обеспечить примерный количественный баланс между подкорпусами разного типа (табл. 1).

Структура корпуса:

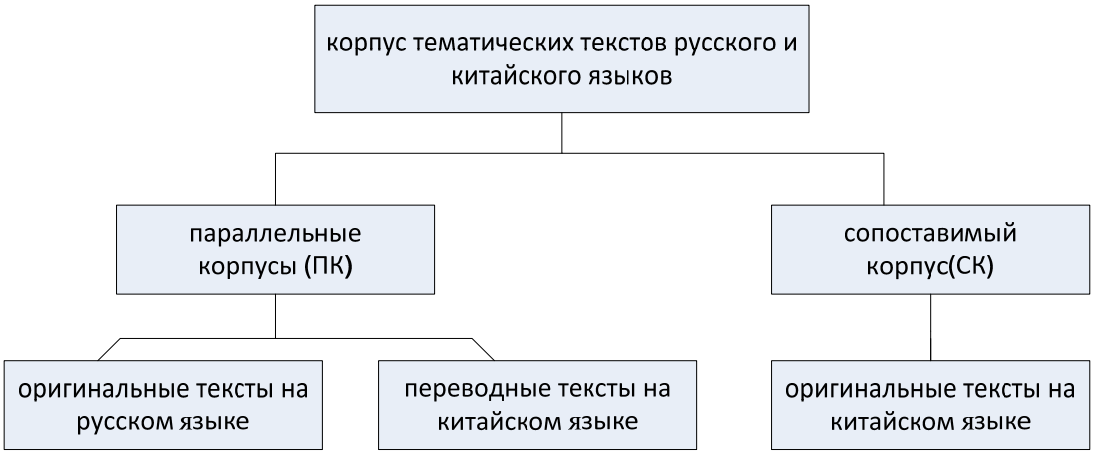


Рис. 1. Структура тематического корпуса русского и китайского языков

Таблица 1

Объем тематических подкорпусов русского и китайского языков

	РС			СС	
	Кол-во слов (русский язык)	Кол-во слов (китайский язык)	Кол-во текстов	Кол-во слов (китайский язык)	Кол-во текстов
Политика и международные отношения	418100	710856	4	303718	4
Лингвистика	568738	855326	3	335546	3
Литературоведение	208643	315258	3	316328	3
Всего	1195581	1891440	10	955582	10

[1] Тексты по переводоведению будут добавлены на следующем этапе

Метаданные

Метаданные тематического корпуса включают в себя следующую информацию: язык, тип текста, автор, переводчик, год издания, год перевода и название монографии. В метаданных четыре типа текстов – политика и международные отношения; лингвистика; литературоведение; переводоведение. Текстовые поля в метаданных повторяют свойство «параллельности» и указываются на двух языках – русском и китайском. Метаданные формируются для обеспечения возможности поиска по динамически формируемым подкорпусам, т. е. пользователь может искать контексты только в интересующих его текстах.

Программное обеспечение

В качестве корпусного менеджера мы используем программу-конкордансер ParaConc [24]. Это программа для создания и эксплуатации параллельных корпусов. С ее помощью можно создавать конкордансы на основе параллельных текстов и выполнять функции по работе с этими корпусами. Одновременно могут быть обработаны до четырех входных корпусов на разных языках. ParaConc изначально поддерживает шрифты для различных языков, включая китайский, японский и корейский. В последней версии конкордансера многоязыковая поддержка обеспечивается за счет кодировки Unicode при условии, что входные тексты представлены в UTF-16. Для построения параллельного конкорданса для двух или более сравниваемых текстов необходимо обеспечить их выравнивание – соответствие по структуре (с точностью до абзацев и предложений). Конкордансер содержит утилиту для полуавтоматического выравнивания текстов по знакам препинания.

В программе реализованы возможности различных режимов поиска. По результатам поиска в окне результатов внутри конкорданса возможен поиск потенциальных переводов. Обеспечивается поиск коллокаций (устойчивых словосочетаний) по различным алгоритмам в диапазоне от одного до четырех слов влево и вправо от ключевого слова. При этом коллокат в строке конкорданса выделяется цветом, что позволяет легко анализировать полученные результаты. Выполняется подсчет частот встречаемости слов и словосочетаний.

Предусмотрена сортировка результатов поиска по трем ключам. Ключом сортировки могут быть и метаданные. Сортировка возможна от начала слова и от конца. Реализованы различные режимы вывода конкорданса, включая вывод в формате HTML. Есть возможность генерации индексов и словников.

Процедуры обработки текстов при создании корпуса

В процессе создания корпуса выполняются следующие операции.

1. Заполнение полей данных и метаданных. Поля, из которых формируются корпус и двуязычный паспорт текста:

id – идентификатор текста;
type – тип текста;
author – автор;
translator – переводчик;
time1 – год издания;
time2 – год перевода;
title – название монографии;
ru – абзац текста на русском языке;
ch – соответствующий ему абзац на китайском языке.

2. Сегментация текста на абзацы и импорт данных в параллельный корпус (рис. 2).

3. Выравнивание текста по предложениям. На первом этапе выравнивание текстов в корпусе по предложениям выполняется автоматически с использованием соответствующей функции ParaConc (точность выравнивания 60-70%) (рис.3).

	id	ru	ch	type	author	translator	time	title
☐	1	УЧЕБНОЕ ПОСОБ	前言/n本书/r旨在/v阐述	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	2	В настоящее в	目前/t，/w至少/d在/p三	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	3	Жестких границ	在/p这儿/r不/d存在/v严	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	4	К примеру, див	比如/v，/w“/w思维/n工	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	5	В литературе н	在/p21/m世纪/n初/f关于	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	6	Характерный п	里/f卡尔德斯/nK的/u畅	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	7	Бизнес -- не в	商业/n不/d是/v战争/n，	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	8	Не случайно д	战略/n经营/vn管理/vn场	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	9	Хотя Бойд и е	尽管w伯伊德/nr及其/c追	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	10	Считается, что	“对于/p所有/b的/u经理	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	11	Вместе с тем о	同时/c显而易见/i的/u是	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	12	Что касается	至于/p谈到/v在/p分析/v	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	13	Шродт, наприме	比如说/l，/w施罗德/nr	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	14	И экономическ	作为/p重要/a成分/n，/w	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	15	Лучшим пример	游戏/n理论/n可以/v作为	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	16	Хотя политоло	尽管/c现在/t政治学/n从	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	17	Вместе с тем н	同时/c必须/d明白/v并/c	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
☐	18	В качестве пр	比如说/l，/w美国/ns议	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

Рис 2. Импорт в базу данных корпуса текстов на русском и китайском языках, выровненных по абзацам

На втором этапе ошибки выравнивания исправляются вручную (100%) (рис. 4).

Предложение русского языка на рис. 3 было переведено двумя предложениями китайского языка, и ParaConc, выравнивающий тексты на двух языках по знакам препинания, сопоставил русскоязычному пред-ложению только первую часть перевода, а второе предложение перевода (в правом нижнем углу на рис. 3) было выровнено со следующим предложением русского текста. Для того чтобы исправить ошибку, второе предложение на китайском языке было вручную перенесено в верхнюю зону (четвертая строка в правой половине на рис. 4). Поэтому выравнивание в автоматическом и в ручном режимах выполняется итерационно, маленькими порциями, с тем чтобы вовремя остановить дальнейшие возможные ошибки выравнивания средствами корпусного менеджера.

4. Загрузка. По завершении выравнивания автоматически средствами ParaConc осуществляется загрузка выровненных текстов в базу данных конкордансера. В интерфейсе конкордансера выбираем последовательно `File – Export – Export CorpusFiles`, и загрузка выполняется (рис. 5).

Рис.3. Автоматическое выравнивание средствами ParaConc

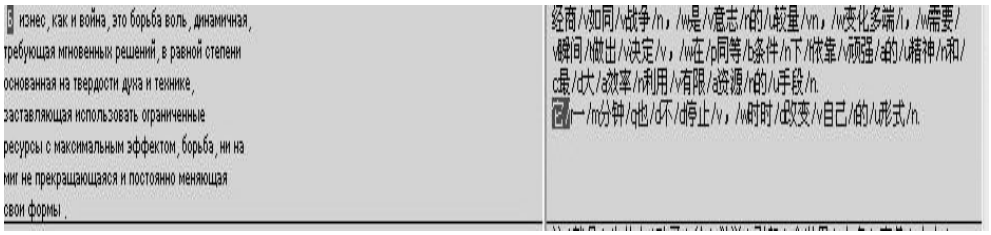


Рис. 4. Исправление ошибок выравнивания

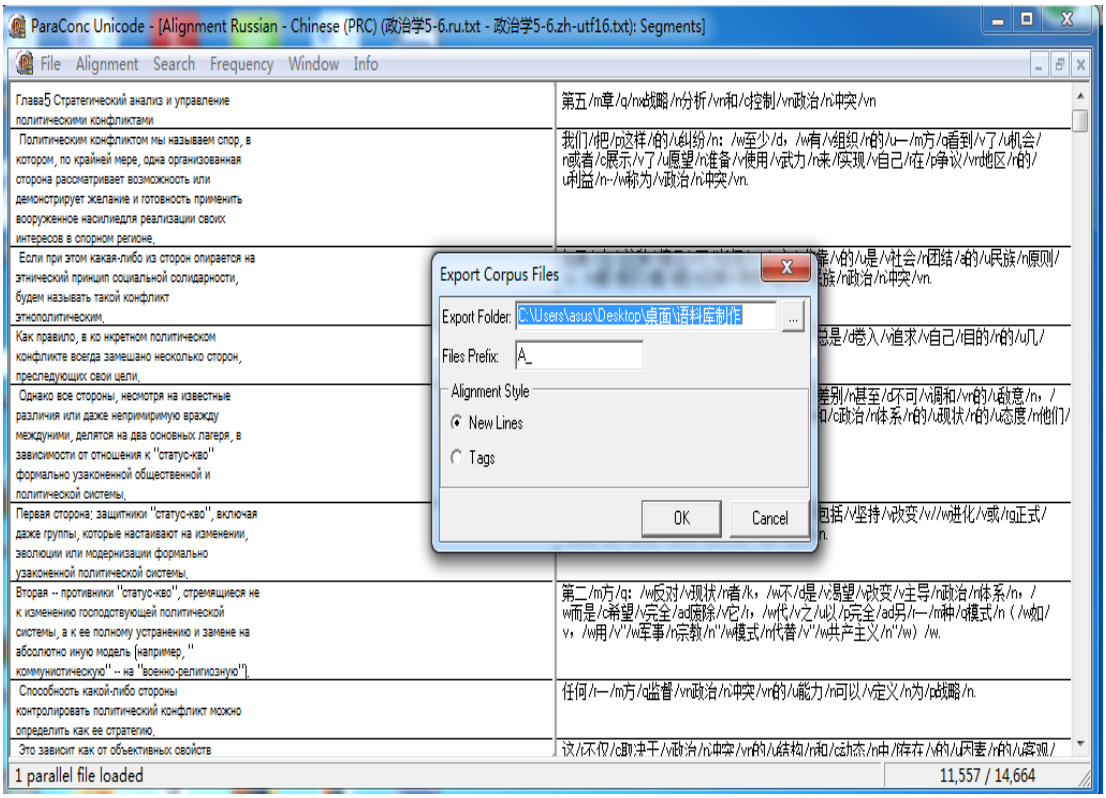


Рис. 5. Автоматическая загрузка текстов в корпус на платформе ParaConc

ПОИСК В КОРПУСЕ

Как уже говорилось, в программе ParaConc реализованы различные режимы поиска: простой поиск по тексту; поиск с использованием регулярных выражений; поиск по метаданным; параллельный поиск; контекстно-обусловленный поиск. На данный момент морфологическая нормализация для русского языка не выполняется. Для того чтобы обеспечить выдачу конкорданса в терминах лексем, поиск осуществляется по словоформам на основе языка регулярных выражений (regular expressions) с возможностью находить все члены словоизменительной парадигмы, что равносильно поиску ключевых слов по леммам.

Язык регулярных выражений RegEx. Языки запросов корпусных менеджеров, представленные в той или иной форме (формализованный язык запросов или графический оконный интерфейс), как правило, базируются на языке регулярных выражений. Регулярные выражения – это строковые записи, задающие правила поиска. Если есть поисковое выражение и какая-либо строка (слово, массив текстов, записи в полях базы данных и т. д.), то операцию проверки, удовлетворяет ли строка поисковому выражению, называют сопоставлением строки и выражения или поиском строк, удовлетворяющих выражению. Если какая-то строка или часть строки успешно сопоставилась с выражением, это называется совпадением (соответствием).

В языке RegEx каждое выражение состоит из одной или нескольких управляющих команд. Некоторые из них можно группировать, и тогда они принимаются за одну команду. Все управляющие команды разбиваются на три класса:

1) *простые символы*, а также управляющие символы, играющие роль их заменителей;

2) *управляющие конструкции* (квантификаторы повторений, оператор альтернативы, группирующие скобки и т. д.);

3) так называемые *мнимые символы* (в строке их нет, но они «помечают» какую-то часть строки – например, ее начало или конец).

Сопоставление может выполняться по простым символам, обозначающим самих себя, по управляющим символам, обозначающим любой символ или класс символов, при этом могут задаваться исключения или альтернативы.

Существует несколько разновидностей языка регулярных выражений. Один из вариантов такого языка реализован в системе ParaConc.

Примеры поиска. Приведём пример поиска с использованием языка регулярных выражений. Для наглядности покажем, как это выглядит на платформе MS Word. Введя регулярное выражение <[Кк]отор*>, мы найдем все члены словоизменительной парадигмы лексемы «который» (рис. 6).

Здесь класс символов [Кк] соответствует любой букве «к», прописной или строчной, а звездочка «*» замещает все возможные формы окончаний.

Такой же режим поиска реализован и на платформе ParaConc. Результатом поиска по запросу <[Кк]отор*> в ParaConc является конкорданс для русского текста (верхняя половина экрана), а внизу помещается соответствующий ему фрагмент текста на китайском языке.

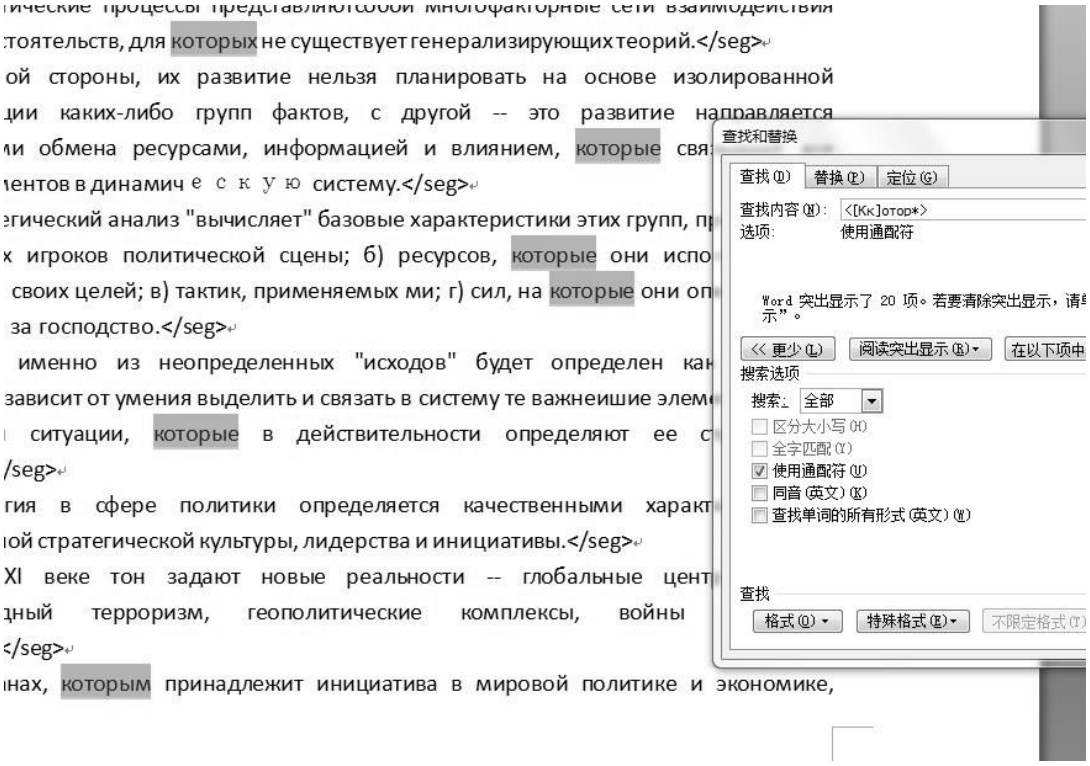


Рис. 6. Результат поиска по выражению <[Кк]отор*> на платформе MS Word

Поиск словосочетаний на китайском языке тоже осуществляется с помощью регулярных выражений. Введя в поисковом окне выражение <一*> (китайский иероглиф «один» плюс звездочка), мы получаем все сочетания с этим иероглифом в китайском подкорпусе в виде конкорданса, а под ним – конкорданс с их русскими эквивалентами.

В настоящее время мы создаем платформу удаленного поиска через Интернет на основе СУБД MySQL. Разработан сайт корпуса, через веб-интерфейс которого реализуется поиск по лексическим единицам с добавлением элементов метаданных. Например, проводя поиск по запросу «Метод анализа иерархий» (等级分析法), мы можем ограничить его типом текста («политика»), или автором («Ожиганов»), или темой («стратегический анализ политики», по-китайски – 政治战略分析, или «стратегия» – 战略), или совокупностью элементов метаданных. Поисковые реквизиты могут вводиться как на китайском, так и на русском языке.

ГЕНЕРАЦИЯ СЛОВНИКА ТЕРМИНОВ

Материал корпуса – специальные тексты гуманитарного направления, и, естественно, в оригиналах и переводах встречается много терминов. Перевод терминов является важнейшей темой переводческих исследований и переводческой практики. Поэтому создание словника или базы данных терминов – одна из задач нашего проекта.

На первом этапе мы ищем термины в текстах на русском языке и их переводы в текстах на китайском языке, вручную выделяем их и автоматически составляем из выровненных терминов новый параллельный текст – двуязычный терминологический словник. При этом термином может быть и словосочетание и одному русскому термину может соответствовать несколько китайских.

На втором этапе мы импортируем выровненный параллельный словник построчно (с помощью инструментов eclipse и java) в базу данных и выполняем генерацию терминологического словника по данным корпуса с возможностью поиска в этой базе. Естественно, такой словник, прежде чем стать словарем, должен пройти интеллектуальную обработку.

ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ НА ОСНОВЕ КОРПУСА

Параллельный корпус переводов с русского языка на китайский позволяет решать различные лингвистические, переводческие и образовательные задачи. Приведем примеры некоторых таких задач.

1. Исследование универсальных принципов перевода на китайский язык на примере анализа перевода неопределенных местоимений и неопределенных наречий, таких как *кто-то, кто-нибудь, кое-кто, что-то, что-нибудь, кое-что, когда-то, когда-нибудь, кое-когда, где-то, где-нибудь, кое-где* и др.

2. Исследование универсальных принципов перевода на китайский язык на основе анализа перевода предикативных наречий, например: *можно, нужно, надо, необходимо, нельзя* и т.д.

3. Исследование универсальных принципов перевода на китайский язык на примере анализа перевода страдательного (пассивного) залога, в том числе страдательных причастий и возвратных глаголов с постфиксом *«-ся»*.

4. Исследование норм перевода сложноподчиненных придаточных предложений с русского на китайский. Например, в процессе перевода сложноподчиненных придаточных предложений с союзом или союзным словом «чтобы» существует четыре вида нормы – экспликация и импликация, упрощение-осложнение, нормализация и отчуждение, с использованием и без использования идиом. Каждый вид нормы проверяется по сравнению с непереводными текстами, т. е. с сопоставимым корпусом.

Приведём пример изучения универсалии осложнения. Сопоставив четыре вида сложноподчиненных придаточных предложений с союзом «чтобы» (придаточные цели, причины, степени/образа действия и изъяснительные) с переводом на китайский язык и на втором шаге – китайские переводы с оригинальными текстами на китайском языке в сопоставимом корпусе, получаем следующий результат (табл. 2).

Таблица 2

Осложнение при переводе придаточных предложений с союзом «чтобы»

Вид придаточных предложений	Кол-во предложений (в переводном тексте)	Кол-во предложений (в непереводном тексте)	Величина осложнения	Величина осложнения (в процентах)
цели	308	289	+19	6,2%
изъяснительные	286	279	+7	2,5%
причины	55	49	+6	10,9%
степени/образа действия	153	140	+13	8,5%
Итого	802	757	+45	В среднем 5,6%

Из табл. 2 видно, что частота осложненных предложений каждого вида в переводных текстах больше, чем в непереводных (сопоставимый корпус). Видимо, переводной китайский язык проявляет нормы осложнения в связи с тем, что в оригинале (русские тексты) сложные предложения используются чаще и грамматика оригинала оказывает влияние на язык перевода.

5. Определение основной переводной единицы на основе корпуса. Мы выдвинули предположение, что пословный перевод на практике не применим. В аспекте прагматики мы всегда имеем дело с пословно-пооборотным переводом. Таким образом, можно утверждать, что словосочетание является основной единицей перевода. В исследовании, проведенном на основе корпуса, мы показали, что значение многих слов определено только в контексте, что окончательно соответствие перевода оригиналу осуществляется только на уровне словосочетания [25].

Например, мы искали в корпусе перевод на китайский язык словосочетаний с предлогом «с», в частности, сочетания «с помощью» (табл. 3).

В корпусе данному словосочетанию соответствуют пять вариантов перевода на китайский язык: «使用、用»; «帮助；借助于»; «用来» и «省去不译». Кроме пятого варианта (пропуск, буквально словосочетание не переведено), остальные четыре являются отдельными независимыми словами и словосочетаниями китайского языка.

6. Практические исследования на основе корпуса перевода дискурсивных маркеров, в том числе, таких как *речь идёт о..., как указывалось, как отмечалось, согласно этому, в соответствии с чем, в результате, следовательно, ввиду этого, в зависимости от этого, в связи с чем, рассмотрим, перейдём к рассмотрению..., итак, короче говоря, иначе говоря, из этого следует...*

7. Практические исследования на основе корпуса правил трансформации предложений с союзами *«который», «так.. как..»* и др.

ВЫВОДЫ

В настоящей работе мы описали цели, значение и процесс создания параллельного корпуса русского и китайского языков, а также способы его использования.

В настоящее время в теории принятия решений, в том числе в научных исследованиях, существует два подхода: нормативный и дескриптивный. Создание описываемого корпуса позволяет анализировать перевод с русского языка на китайский в аспекте дескриптивного подхода, на основе реального языкового материала. На платформе корпуса возможно проведение не только качественного, но и количественного исследования, не только оценка переводных текстов (их достоинства и недостатки), но и исследования природы и универсальности переводного языка.

Таблица 3

Примеры переводов словосочетания «с помощью» в корпусе

原文	译文
...остижения превосходства одной из сторон [[с помощью]] применения вооруженной силы.	……一/m方/q如何/r使用/v武装力量/l来/f取得/v优势/n的/u原则/n和/c方式/n.
...ло -- выражение математической формулы, [[с помощью]] которой Парето доказывает, что распреде ...	这个/r规则/n是/v个/q数学/n公式/n，/w巴莱多/nr用/p它/r证明/v了/u社会/n财富/n和/c收入/n的/u分配……
... азывают как угодно. Метод, [[с помощью]] которого приобретают знания об этих отн ...	帮助/v获得/v这些/r关系/n的/u知识/n的/u方法/n具有/v不大/d的/u意义/n.
... в качестве самостоятельных переменных, [[с помощью]] которых описывается и объясняется полит ...	……作为/v独立/vd变化/v起/v作用/n的/u社会关系/l的/u一定/b总和/n，/w政治/n行为/n借助于/v这些/r关系/n而/c被/p描述/v、/w诠释/v.
... ться “управлять” иерархической моделью [[с помощью]] математических средств, что помогает ем ...	……他/r可以/v尝试/v用/p帮助/v他/r找到/v正确/ad解决/v问题/n方法/n的/u数学/n手段/n来/f“/w操纵/v”/w等级/n模型/n.

Способы разработки параллельного корпуса еще несовершенны, и мы будем их улучшать. В частности, намечено разработать дополнительные программы предварительной обработки и разметки текстов. Также планируется автоматическая лемматизация текстов русскоязычной части корпуса. Предстоит разработать дополнительные средства в части обеспечения гибкого управления поиском и сервиса в подсистеме выдачи результатов. В подсистеме генерации терминологических словарей на основе корпуса намечена разработка модуля автоматического выявления терминологической лексики на основе различных статистических методов. Решение этих и других задач позволит оптимизировать последующую работу по пополнению параллельного корпуса и по его использованию, станет моделью для разработки следующих корпусов.

СПИСОК ЛИТЕРАТУРЫ

1. Languages in contrast: Papers from a Symposium on Text-based Cross-linguistic Studies / eds. K. Aijmer, B. Altenberg, M. Johansson. – Lund: Lund University Press, 1996.
2. Aston G.. Corpus use and learning to translate // Textus. – 1999. – № 12. – P. 289-314.
3. Zanettin F. Bilingual comparable corpora and the training of translators // Meta: Translators' Journal. – 1998. – Vol. 43, № 4. – P. 616-630.
4. Kaeding F. W. Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsauss-chuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. – Zweiter Teil: Buchstaben-zählungen. – Steglitz bei Berlin: Selbstverlag des Herausgebers, 1897-1898.
5. Захаров В.П. Корпусная лингвистика. – СПб.: СПбГУ, 2005.
6. Laviosa S. Corpora and translation: the methods and theories of corpus work in translation. – Manchester: 2003.
7. Granger S. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies? – Louvain: University of Louvain, 2003.
8. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. 2-е изд., перераб. и дополн. – СПб.: СПбГУ, 2013.
9. Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов // Корпусная лингвистика–2004: Труды международной конференции. – СПб.: СПбГУ, 2004. – С. 55-64.
10. Zhan W. D., Chang B. B., Duan H. M., Zhang H. R. Recent developments in Chinese corpus research // The 13th NIJL International Symposium, Language Corpora: Their Complication and Application. – №3. – Tokyo, 2006.
11. McEnery A., Xiao Z. H. Aspect making in English and Chinese: Using the Lancaster corpus of Mandarin Chinese for contrastive language study // Literary and Linguistic Computing. – 2003. – № 4. – P. 361-378.
12. Liu Z. Q., Tian L., Liu C. <Hongloumeng> zhongyingwen pingxing yuliaoku de chuangjian // Dangdai yuyanxue. – 2008. – № 4. – P. 329-339.
13. Cao D. F. Rihan pingxing yuliaoku de sheji yu jianshe // waiyu jiaoxue yu yanjiu – 2006. – № 3. – P. 221-227.
14. Bernardini S., Zanettin F. When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals // Translation universals: Do they exist? / eds. A. Mau-ranen, P. Kujamaki. – Amsterdam: John Benjamins, 2004. – P. 51-62.
15. Cui W., Zhang L. Ehan fanyi pingxing yuliaoku jiqi yingyong yanjiu // Jiefangjun waiguoyu xueyuan xuebao. – 2014. – № 1. – P. 81-87.
16. Ma Q. Z. Zhuming zhongnian yuyan xuejia zixuanji. – Ma Qingzhu juan:. Anhui jiaoyu chubanshe, 2002.
17. Zhao M. S. E hanyu duibi yanjiu. – Shanghai yiwen chubanshe, 1994.
18. Lu J. M. Ci de juti yiyi dui juzi yisi lijie de yingxiang // Hanyu xuexi. – 2004. – № 2. – P. 1-5.
19. Zhao S. J. Shi lun cihui yuyi dui yufa de jueding zuoyong // Wuhan daxue xuebao. – 2008. – № 2. – P. 173-179.
20. Baker M. Corpus linguistics and translation studies: Implications and applications // Text and Technology: In Honour of John Sinclair / eds. F. Baker, Tognini-Bonelli. – Amsterdam/Philadelphia: John Benjamins, 1993. – P. 233-250.
21. Baker M. A corpus-based view of similarity and difference in translation // International Journal of Corpus Linguistics. – 2004. – № 2. – P. 167-193.
22. Hu X. Y. Yuliaoku fanyi yanjiu yu fanyi pubianxing // Shanghai keji fanyi. – 2004. – № 4. – P. 47-49.
23. Huang C. N. Yuliaoku yuyanxue // Shangwu yinshuguan. – Bejing, 2004.
24. Barlow M. ParaConc: Concordance Software for Multilingual Parallel Corpora // Proceedings of the Third International Conference on Language Resources and Evaluation. LREC Workshop № 8: Workshop on Language Resources in Translation Work and Research. – 2002. – P. 20-24.
25. Tao Y. Jiyu ehan pingxing yuliaoku de fanyi danwei yanjiu // Waiyu jiaoxue. – 2015. – № 1. – P. 108-113.

ПРИЛОЖЕНИЕ 1

Состав параллельного корпуса

№ п/п	Название книги	Автор	Год издания	Издательство	Переводчик
1.	Стратегический анализ политики	Ожиганов Э.Н.	2006	Аспект-пресс	Ху Гумин и др.
2.	Экополитология и глобалистика	Костин А.И.	2005	Аспект-пресс	Ху Гумин и др.
3.	Мировая политика	Лебедева М.М.	2007	Аспект-пресс	Лю Цайци и др.
4.	Социология международных отношений	Цыганков А.П.	2006	Аспект-пресс	Лю Цайци и др.
5.	Интегральное описание языка и системная лексикография (1)	Апресян Ю.Д.	1995	Языки славянской культуры	Ду Гуйси
6.	Интегральное описание языка и системная лексикография (2)	Апресян Ю.Д.	1995	Язык славянской культуры	Ду Гуйчжи
7.	Семантика предложения и нереферентные слова	Шатуновский И.Б.	1996	Языки славянской культуры	Сюй Энькуй
8.	Автор и герой в эстетической деятельности: проблема отношения автора к герою	Бахтин М.М.	2000	Азбука	Ли Хуйфань и др.
9.	Творчество Франсуа Рабле и народная культура Средневековья и Ренессанса	Бахтин М.М.	1990	Художественная литература	Ли Чжаолинь и др.
10.	Проблемы поэтики Достоевского	Бахтин М.М.	1963	Советский писатель	Лю Ху

ПРИЛОЖЕНИЕ 2

Состав сопоставимого корпуса

№ п/п	Название книги	Автор	Год издания	Издательство
1.	Расширение НАТО на восток и стратегический выбор России	Лю Цзюнь, Ли Хайдун	2010	Хуадуншифаньдасюе
2.	Современная Россия: процесс политического развития и выбор внешней стратегии	Фань Цзаньцун	2004	Шиши
3.	Политическая мысль современной России	Чжан Шухуа, Лю Сяньчжун	2003	Синьхуа
4.	Взаимоотношение ЕС и РФ	Луо Чжиган	2009	Чжунгуошехуйкесюе
5.	Современная русская семантика	Чжан Цзяхуа	2003	Шаньу
6.	Развитие и тенденция исследования русского языка в конце 20 века	Ду Гуйчжи	2000	Шуодушифаньдасюе
7.	Количественное исследование лексикологии	Сюй Энькуй	2006	Хэлунцзян

№ п/п	Название книги	Автор	Год издания	Издательство
8.	Современная русская литература	Хоу Вэйхун	2013	Чжунгуошехуйкесюе
9.	Стилистика и перевод романов	Ху Гумин	2004	Шанхайивэнь
10.	Бердяев и русская литература	Гэн Хайин	2009	Шанхайшудянь

Материал поступил в редакцию 20.01.15

Сведения об авторах

ТАО Юань – доктор филологических наук, доцент Факультета русского языка Шэньсийского педагогического университета
E-mail: tao1973@mail.ru

ЗАХАРОВ Виктор Павлович – кандидат филологических наук, доцент кафедры математической лингвистики Санкт-Петербургского университета
E-mail: vz1311@yandex.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн. документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ – http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно и без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе:

- **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:
  - **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы.
  - **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков).
  - **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
  - **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
  - Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.
- **в коммуникативных форматах** iso-2709, мекоф, txt на любых видах электронных носителей.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Административная группа БД ВИНИТИ – 8-499-155-45-01,
8-499-155-45-02
Отдел взаимодействия с потребителями – 8-499-155-45-25,
8-499-155-46-20
E-mail: davydova@viniti.ru , csbd@viniti.ru
WWW: http://www.viniti.ru FAX – 8-499-155-45-01, 8-499-155-45-25

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефон: 8 (499) 155-46-20

Телефон/Факс: 8 (499) 155-45-25

E-mail: zinovyeva@viniti.ru, davydova@viniti.ru

Федеральное государственное бюджетное учреждение науки
ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

«Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, широкое тиражирование которых, как правило, в силу их узкой специализации, не считается целесообразным, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях».

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2013).

Депонированные научные работы находятся на хранении в депозитарии ВИНИТИ РАН, копии работ предоставляются заинтересованным организациям и специалистам на бумажном и электронном носителях и являются официальной публикацией.

Информация о депонированных научных работах включается в информационные издания ВИНИТИ РАН, в РЖ ВИНИТИ РАН и БД ВИНИТИ РАН и аннотированный библиографический указатель «Депонированные научные работы».

Подать научную работу на депонирование можно, обратившись в Отдел депонирования ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Отдел депонирования научных работ.
Тел.: 8 (499) 155-43-28, Факс: 8 (499) 943-00-60.
e-mail: dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.ru**