АКАДЕМИЧЕСКАЯ ЖИЗНЬ

УДК 004:94

РОССИЙСКО-ФРАНЦУЗСКИЙ СЕМИНАР «ТЕКСТОМЕТРИЯ И КОРПУСЫ РУССКИХ ТЕКСТОВ»[1]

Д. А. Гагарина
Пермский государственный национальный исследовательский университет, 614990, Пермь, ул. Букирева, 15
dinara@psu.ru

Дается обзор проектов, представленных на российско-французском семинаре «TXM и корпусы русских текстов», прошедшем в Лионе в июне 2015 г. Обсуждаются теоретические и прикладные проблемы создания информационных ресурсов в области истории, политологии и лингвистики на основе российских и французских исторических источников. Описаны возможности платформы TXM (TeXtoMetrie) для исследований в области цифровой гуманитаристики (digital humanities), среди которых инструменты анализа текстов, построение конкордансов, поиск мотивов, частотные словари, различные статистические функции, классификации, анализ совместной встречаемости.

Ключевые слова: digital history, digital humanities, историческая информатика, информационные ресурсы, историко-ориентированные системы, компьютерная лингвистика.

С 8 по 12 июня 2015 г. в Лионе состоялся российско-французский семинар «TXM и корпусы русских текстов», посвященный обсуждению теоретических и прикладных проблем создания современных информационных ресурсов для научных исследований в области истории, политологии, лингвистики и других гуманитарных наук.

Семинар организован Высшей нормальной школой Лиона (ENS de Lyon), лабораторией ICAR (Interactions, Corpus, Apprentissages, Representations), лабораторией Labex ASLAN (Advanced Studies on LANguage complexity) совместно с Библиотекой Дени Дидро. Инициатором и руководителем семинара выступил профессор, доктор наук, научный сотрудник лаборатории ICAR Национального центра научных исследований (CNRS) и Лионского университета Алексей Лаврентьев.

Центральное мероприятие в программе – исследовательский семинар, который прошел в Высшей нормальной школе Лиона. С французской стороны в семинаре приняли участие Кристин Бое, Анн Мэтр, Катрин Сеньере (Библиотека Дени Дидро, Лион), Серж Эйден, Бенедикт Пинсеми (лаборатория ICAR), Мишель Тиссье (Университет Рена) и др., с российской стороны – сотрудники Пермского государственного национального исследовательского университета: профессор Сергей Корниенко, профессор Игорь Кирьянов, доцент Динара Гагарина, старший преподаватель Надежда Поврозник, а также доцент Белгородского университета Елена Маркова. На семинаре были представлены и обсуждены проекты создания научных информационных ресурсов на основе российских и французских исторических источников.

Открыла семинар К. Бое (Chr. Boyer), директор Библиотеки Дени Дидро. Во вступительном слове она отметила значимость проектов создания цифровых корпусов исторических источников и необходимость объединения усилий в этой области.

А. Мэтр (A. Maître), М. Тиссье (M. Tissier) и К. Сеньере (C. Seigneret) представили доклад «Корпус "Борис Чичерин"».

«Корпус "Борис Чичерин" – междисциплинарный проект, создаваемый совместно группой исследователей и библиотекой под руководством профессора С. Мартен (S. Martin), специалиста в области истории России и русского языка, вице-президента по учебной работе Высшей нормальной школы. Целью проекта является изучение либеральной мысли в России на основе произведений Бориса Николаевича Чичерина (1828–1904), историка, философа и теоретика русского либерализма [*Martin*, 2013].

Проект включает публикации произведений на русском языке из коллекции Славянского фонда библиотеки Дени Дидро, дополненных источниками из ряда других библиотек Франции. Корпус состоит из книг и статей Б.Н. Чичерина и некоторых текстов его современников. Ожидае-

мый объем корпуса – 13000–14000 страниц. Тексты будут свободно доступны в Интернете в формате PDF на сайте Библиотеки Дени Дидро [Bibliotheque Diderot…] в соответствующем разделе.

А. Мэтр, руководитель отдела сохранения наследия Славянского фонда Библиотеки Дени Дидро, рассказала о целях проекта, принципах и этапах его реализации, составе исполнителей, отборе тестов для включения в корпус. Кроме того, она сообщила о коллекции Славянского фонда и истории ее формирования.

М. Тисье представил возможности использования корпуса «Борис Чичерин» в исторических исследованиях, для изучения истории либерализма и других вопросов. На материале серии статей «Опыты по истории русского права», напечатанных в журнале «Русский Вестник» и газете «Московские ведомости» в 1856–1857 гг., был показан процесс создания хронологического и биографического справочников, глоссария, коллекции цитат, а также были определены особенности разметки текстов, их орфографической нормализации и представления с помощью платформы TXM.

Доклад К. Сеньере был посвящен техническим и методическим аспектам работы над проектом «Корпус "Борис Чичерин"», особенностям оцифровки источников, программных средств для распознавания, метаданных, форматов и принципов хранения файлов электронных версий источников.

Е.И. Маркова и А.М. Лаврентьев предложили доклад «TXM и цифровая филология: презентация проекта мультиуровневого издания средневекового трактата о здоровье». В докладе описаны методика создания электронной версии издания медицинского трактата «L'enseignement ou la manière de garder et conserver la santé» («Трактат о сохранении здоровья»), представляющего собой перевод на среднефранцузский язык латинского произведения Гвидо Парато «Libellus de sanitate conservanda» (1459) [*Lavrentiev, Markova,* 2014]., а также методологические и технологические решения подготовки издания, решение задачи орфографической нормализации текста, принципы разметки источника.

Опыт реализации пермских проектов создания историко-ориентированных информационных систем был представлен в докладах Д.А Гагариной, И.К. Кирьянова, С.И. Корниенко и Н.Г. Поврозник.

Доклад И.К. Кирьянова был посвящен серии проектов в области истории парламентаризма и интернет-порталу «Парламентская история позднеимперской России» [*Гагарина, Кирьянов, Корниенко,* 2011; *Кирьянов, Корниенко, Гагарина, Рябухин,* 2010; *Кирьянов, Корниенко, Гагарина,* 2014]. В докладе представлены модели созданных информационных систем, структура интернет-портала, показаны возможности их использования для изучения становления и эволюции Государственной думы Российской империи, ее деятельности и депутатского корпуса. Созданные ресурсы не только позволяют решить задачи сохранения, организации и визуализации основных исторических источников по истории парламентаризма (Стенографические отчеты Государственной думы и Государственного совета, указатели к ним и др.), но и содержат инструментарий для поддержки научных исследований по теме.

С.И. Корниенко выступил с докладом «Историко-ориентированные системы на основе губернской периодики: пермские проекты». В докладе рассмотрена серия проектов по созданию информационных систем на основе пермской губернской периодики, раскрыты теоретические и прикладные проблемы создания и использования указанного вида ресурсов. Особое внимание уделено сохранению и изучению пермских газет периода Первой мировой войны [*Корниенко, Гагарина, Митина,* 2014] и Гражданской войны (периода оккупации Перми войсками Колчака) [*Корниенко, Гагарина, Масленников, Пигалева,* 2013]. Были также обсуждены вопросы оцифровки и распознавания газетной периодики, разработки информационных моделей различного типа периодических изданий, проблемы их описания, форматы разметки и представления текстов. В рамках доклада были продемонстрированы разработанные ресурсы и возможности их использования для сохранения историко-культурного наследия и научных исследованиях.

Д.А. Гагарина выступила с докладом «Историко-ориентированные информационные системы» («System of (history-oriented information) systems»), посвященным проекту «Историко-ориентированные информационные системы: методологические, теоретические и прикладные проблемы создания и использования» [*Корниенко, Гагарина,* 2014] и созданному в рамках проекта сайту [Историко-ориентированные…]. Ресурс включает каталог историко-ориентированных систем и каталог публикаций о них. Каждая система и публикация имеют подробное описание, структура

которого разработана в рамках проекта. В ходе доклада продемонстрированы поисковый инструментарий и другие возможности ресурса, а также обсуждены проблемы описания и каталогизация историко-ориентированных ресурсов.

В докладе Н.Г. Поврозник «Информационная система "Журналы земских собраний как источник изучения истории местного самоуправления в России (II половина XIX – начало XX века)"» представлены теоретико-методологические и практические аспекты разработки информационной системы для сохранения и анализа журналов губернских земских собраний, продемонстрированы интерфейс информационной системы, шаблоны для ввода данных и вывода результатов поиска [*Корниенко, Масленников, Шабалина*, 2005]. Разработанный ресурс позволяет получать информацию из земских журналов как единого текста в рамках всего их массива, погубернских и поуездных его долей на протяжении всего периода земской истории и ее отдельных этапов. На основе системы проведены исследования, посвященные различным вопросам земской истории.

Семинар предусматривал изучение и овладение практическими навыками работы с платформой TXM, успешно применяемой для создания корпусов текстов исторических и историографических источников и их качественного анализа с помощью компьютерных методов. В рамках семинара был проведен тренинг, включающий лекционные и практические занятия по технологии XML-TEI и текстометрическому анализу на платформе TXM.

Платформа TXM [TXM…] может использоваться специалистами по разным гуманитарным наукам (истории, политологии, лингвистике, литературоведению и др.), работающими с корпусами текстов. TXM предлагает полный набор инструментов анализа текстов (построение конкордансов, поиск мотивов, частотные словари и т.д.), основанных на поисковой машине CQP [CWB…], и большое число статистических функций (факторный анализ соответствий, классификация, анализ совместной встречаемости и т.д.), основанных на пакетах R [The R Project…].

TXM разрабатывается в лаборатории ICAR [ICAR], является свободно распространяемым программным обеспечением с открытым кодом. Оно включает среду для анализа текстов и корпусов в различных форматах, кроме того может использоваться в виде веб-приложения он-лайн. С 2012 г. TXM обладает русскоязычным интерфейсом. Платформа позволяет импортировать и проводить автоматическую морфологическую разметку и лемматизацию текстов на различных языках, в том числе на современном русском языке.

В ходе семинара С. Эйденом была прочитана лекция «Введение в текстометрию и платформу TXM», А. Лаврентьев провел практические занятия по темам: Функции поиска и качественного анализа; Функции количественного анализа; Подготовка и импортирование корпуса.

Одним из мероприятий семинара стала экскурсия в Славянский фонд Библиотеки Дени Дидро, основу которого составляет коллекция католического священника-иезуита князя Ивана Гагарина. В ходе экскурсии участники семинара познакомились с процессами оцифровки, обсудили роль библиотек в создании научных электронных ресурсов и формировании современной информационной среды гуманитарных наук.

На заключительном этапе семинара состоялся круглый стол, на котором обсуждались перспективы совместных проектов, обмена данными и инструментами анализа.

Прошедший семинар продолжил серию научных мероприятий, организованных лабораторией исторической и политической информатики Пермского университета совместно с европейскими коллегами. Так, в 2009 г. в Университете Граца (Австрия) был проведен совместный российско-австрийский семинар «Документирование и анализ историко-культурного наследия методами исторической информатики», соруководителями семинара стали С.И. Корниенко и И. Кропач [*Горбачева, Гагарина, Сметанин*, 2009]. По материалам докладов семинара издан сборник трудов «Documentation and Analysis of the Historical and Cultural Heritage by Historical Information Science Methods» [Documentation and Analysis…, 2009]. В 2012 г. в Вене (Австрия) состоялся семинар «Компьютерные методы сохранения и изучения объектов культурного наследия», организованный совместно с лабораторией компьютерного зрения Венского технологического университета, руководители семинара – С.И. Корниенко и Р. Саблатниг. В рамках подобных семинаров проходит обмен опытом между учеными Пермского и европейских университетов, знакомство с работой библиотек, архивов и специализированных лабораторий, достигаются договоренности по реализации совместных исследовательских, издательских и учебных проектов.

Примечания

[1] Статья выполнена при поддержке РФФИ, грант № 13-06-00655 «Историко-ориентированные информационные системы: методологические, теоретические и прикладные проблемы создания и использования».

Библиографический список

Bibliotheque Diderot de Lyon. URL: http://bibliotheque-diderot.org/.
CWB. The IMS Open Corpus Workbench. URL: http://cwb.sourceforge.net (дата обращения: 01.08.2015).
Documentation and Analysis of the Historical and Cultural Heritage by Historical Information Science Methods: Proceedings of the Joint Seminar (held at Graz, April, 15–17, 2009). Series of the Institute of History (University of Graz). Vol. 18 / eds. by S. I. Kornienko and I. H. Kropač; Perm University. Perm; Graz, 2009. 156 p.
ICAR. URL: http://icar.univ-lyon2.fr
Lavrentiev A., Markova E. From the Holy Grail to the Good Health: a Digital Edition of a 15th Century French Medical Treatise on the BFM Web Portal // Textual Heritage and Information Technologies. El'Manuscript-2014 / Cyrillo-Methodian Research Center and Izhevsk University. Varna, 2014. P.164–166.
Martin S. Des usages de la liberté: abolition du servage et paysannerie chez Boris Tchitchérine // ILCEA. 2013. № 17. URL : http://ilcea.revues.org/1784 (дата обращения: 01.08.2015).
The R Project for Statistical Computing. URL: http://www.r-project.org (дата обращения: 01.08.2015).
TXM. Unicode-XML-TEI text/corpus analysis platform. URL: https://sourceforge.net/projects/txm (дата обращения: 01.08.2015).
Гагарина Д.А., Кирьянов И.К., Корниенко С.И. Историко-ориентированные информационные системы: опыт реализации «пермских» проектов // Вестник Пермского университета. Сер.: История. 2011. № 2 (16). С. 35–39.
Горбачева Н.Г., Гагарина Д.А., Сметанин А.В. Российско-австрийский научный семинар «Документирование и анализ историко-культурного наследия методами исторической информатики», Грац, 15–17 апреля 2009 г. // Вестник Пермского университета. Сер.: История. 2009. Вып. 2(9); Сер.: Политология. Вып. 2(6). С. 136–142.
Историко-ориентированные информационные системы. URL: http://digitalhistory.ru/ (дата обращения: 01.08.2015).
Кирьянов И.К., Корниенко С.И., Гагарина Д.А. Интернет-портал «Парламентская история позднеимперской России»: возможности поддержки научных исследований // Информ. бюл. Ассоциации «История и компьютер». № 42. 2014. С. 27–29.
Кирьянов И.К., Корниенко С.И., Гагарина Д.А., Рябухин И.В. Информационный ресурс по парламентской истории России начала XX в. // Власть. 2010. № 12. С. 83–86.
Корниенко С.И., Гагарина Д.А. Электронный каталог историко-ориентированных информационных систем // Информ. бюл. Ассоциации «История и компьютер». 2014. № 42. С. 115–116.
Корниенко С.И., Гагарина Д.А., Масленников Н.Н., Пигалева С.В. Источнико-ориентированная база данных как основа информационной системы для сохранения и изучения пермских газет колчаковского периода // Круг идей: базы данных в исторических исследованиях. Барнаул, 2013.С. 140–155.
Корниенко С.И., Гагарина Д.А., Митина Р.В. Пермская губернская периодика Первой мировой войны: создание информационной системы // Информ. бюл. Ассоциации «История и компьютер». 2014. № 42. С. 218–219.
Корниенко С.И., Масленников Н.Н., Шабалина Д.В. Журналы земских собраний: проблемы создания информационной системы // Круг идей: алгоритмы и технологии исторической информатики. М.; Барнаул, 2005. С. 153–165.

Дата поступления рукописи в редакцию 07.08.2015

RUSSIAN-FRENCH SEMINAR «TEXTOMETRY AND CORPUSES OF RUSSIAN TEXTS»

D. A. Gagarina

Perm State University, Bukirev str., 15, 614990, Perm, Russia
dinara@psu.ru

The article presents a review of the Russian-French seminar «TXM and corpuses of Russian texts» held at Lyon, on June, 8th-15th, 2015. The seminar was devoted to the discussion of theoretical and applied problems of creation of modern information resources for researches in History, Political Science, Linguistics and other Humanities. The seminar was organized by the Ecole Normale Superieure of Lyon, Laboratory ICAR, Labex ASLAN Laboratory in conjunction with Denis Diderot Library. The article provides a review of the seminar program and reports. The central event of the program was a research seminar held in the Ecole Normale Supérieure of Lyon. Projects of creation of academic electronic resources based on Russian and French historical sources were presented and discussed at the seminar. Among them, there were «Corpus Boris Chicherin», digital edition of a 15th-century French medical treatise, internet portal «Parliamentary History of Pre-Revolutionary Russia», «Zemstvo journals», as well as the series of information systems based on province periodicals. The second part of the seminar was devoted to the TXM workshop. The article describes the TXM opportunities for researches in Digital Humanities.

Key words: Digital History, Digital Humanities, Historical Information Science, information resources, history-oriented information systems, Computer Linguistics.

References

Bibliotheque Diderot de Lyon. URL: http://bibliotheque-diderot.org/.

CWB. The IMS Open Corpus Workbench. URL: http://cwb.sourceforge.net.

Documentation and Analysis of the Historical and Cultural Heritage by Historical Information Science Methods: Proceedings of the Joint Seminar (held at Graz, April, 15–17, 2009). Series of the Institute of History (University of Graz), vol. 18 / eds. by S. I. Kornienko and I. H. Kropač. Perm; Graz, 2009.

ICAR. URL: http://icar.univ-lyon2.fr.

Lavrentiev A., Markova E. From the Holy Grail to the Good Health: a Digital Edition of a 15th Century French Medical Treatise on the BFM Web Portal. *Textual Heritage and Information Technologies. El'Manuscript-2014.* Varna, Bulgaria: Cyrillo-Methodian Research Center and Izhevsk University, 2014.

Martin S. Des usages de la liberté: abolition du servage et paysannerie chez Boris Tchitchérine. *ILCEA.* 2013. No 17.

The R Project for Statistical Computing. URL: http://www.r-project.org.

TXM. Unicode-XML-TEI text/corpus analysis platform. URL: https://sourceforge.net/projects/txm.

Gagarina D.A., Kiryanov I.K., Kornienko S.I. Istoriko-orintirovannye informatsinnye sistemy: opyt realizatsii permskikh proektov. *Vestnik Permkogo universiteta. Seriya Istoriya.* 2011. No. 2 (16).

Gorbacheva N.G., Gagarina D.A., Сметанин А.В. Rossisko-avstriskiy nauchniy seminar «Documentirovanie i analiz istoriko-kulturnogo naslediya metodami istorocheskoi informatiki», Grats, 15-17 aprelya 2009. *Vestnik Permkogo universiteta. Seriya Istoriya.* 2009. No. 2(9). *Seriya Politologiya.* 2009. No. 2(6). Istoriko-orintirovannye informatsinnye sistemy. URL: http://digitalhistory.ru/.

Kiryanov I.K., Kornienko S.I., Gagarina D.A. Internet-portal «Parlametskaya istoriya pozdneimperskoi Rossii»: vozmozhnosti podderzhki nauchnykh issledovanii. *Informatsioniy bulleten Assotsiatsii «Istoriya i komputer».* 2014. No. 42.

Kiryanov I.K., Kornienko S.I., Gagarina D.A., Ryabukhin I.V. Informatsionniy resurs po parlametskoi istorii Rossii nachala XX v. *Vlast.* 2010. No. 12.

Kornienko S.I., Gagarina D.A. Elektroniy katalog istoriko-orintirovannykh informatsinnykh system. *Informatsioniy bulleten Assotsiatsii «Istoriya i komputer».* 2014. No. 42.

Kornienko S.I., Gagarina D.A., Maslennikov N.N., Pigaleva S.V. Istoriko-orintirovannaya baza dannykh kak osnova informatsinnoi sistemy dlya sokhraneniya i izucheniya permskikh gazet kalchakovskogo perioda. *Krug idei: bazy dannykh v istoricheskikh issledovaniykh.* Barnaul, 2013.

Kornienko S.I., Gagarina D.A., Mitina R.V. Permskaya gubernskaya periodika Pervoi mirovoi voiny: sozdanie informatsionnoi systemy. *Informatsioniy bulleten Assotsiatsii «Istoriya i komputer».* 2014. No. 42.

Kornienko S.I., Maslennikov N.N., Shabalina D.V. Zhurnaly zemskikh sobrnii: problem sozdaniya informatsionnoi systemy // Krug idei: algoritmy i technologii istoricheskoi informatiki. Moskva; Barnul, 2005.