Влияние корпусных технологий на развитие диахронической лингвистики: пример Франции

Alexei Lavrentiev

Влияние корпусных технологий на развитие диахронической лингвистики: пример Франции

Лаврентьев Алексей Михайлович
Лаборатория ICAR, Labex ASLAN, Национальный центр научных исследований (CNRS), Лион, Франция, alexei.lavrentev@ens-lyon.fr

Ключевые слова: диахрония, французский язык, историческая грамматика, средневековая литература, текстометрия
Keywords: diachrony, French language, historical grammar, medieval literature, reference corpora, textometry

Title: Influence of corpus technologies on the development of diachronical linguistics: the case of France.

Summary:
This paper presents an overview of the development of corpora for research on the history of the French language and discusses some methodological problems related to the creation and use of such corpora. These include the sustainability and the accuracy of data as well copyright and open access issues. Examples of corpus-based research are given to demonstrate the relevance of corpus data.

1. Из истории корпусных исследований во Франции

Исследования в области истории французского языка, опирающиеся на материал корпуса текстов, проводились еще задолго до появления современных компьютерных технологий. В 1875 г. Ж.-Ж. Лекультр [Lecoultre] публикует работу «О порядке слов у Кретьена де Труа», основанную на систематическом анализе произведений этого автора. В 1919 г. выходит в свет «Маленький синтаксис» Л. Фуле [Foulet], построенный на данных корпуса из 9 текстов конца XII – первой половины XIII в., подобранных с учетом разнообразия жанрового состава, диалектных особенностей языка и социального статуса авторов.

В то же время большинство исследований носят «импрессионистский» характер: примеры из текстов используются как иллюстративный материал, а выводы делаются из наблюдений авторов и данных «вторичной» литературы (комментариев публикаторов текстов).

В 1960-е гг. по инициативе П. Эмбса (P. Imbs) в г. Нанси открывается Центр исследований тезауруса французского языка (CRTLF, в настоящее время «наследником» этого центра является лаборатория ATILF, http://www.atilf.fr), в котором начинается работа по созданию электронной базы текстов для лексикографического использования (создания словаря-тезауруса французского языка, ныне обозначаемого аббревиатурой TLF). Со временем эта база текстов, охватывающая период с XVI по XX в. превратится в ресурс Frantext, широко используемый французскими лингвистами и литературоведами вплоть до нашего времени.

В 1970-е – начале 1980-х гг., однако, использование электронных ресурсов лингвистами было весьма ограниченным ввиду их дороговизны и технической сложности, а тексты, датируемые до 1500 г., просто не были оцифрованы.

В 1979 году К. Маркелло-Низья [Marchello-Nizia] публикует «Историю французского языка в XIV – XV веках», явившуюся результатом кропотливого «ручного» анализа корпуса из 43 среднефранцузских текстов. В 1980-е годы в той же

лаборатории, где создавался Frantext, под руководством Р. Мартена начинается работа над базой текстов для словаря среднефранцузского языка (DMF), а в 1989 г. К Маркелло-Низья инициирует проект Базы средневекового французского языка (BFM), которая была призвана дополнить разрабатываемые в Нанси ресурсы текстами старейшего периода истории французского языка (IX – XIII вв.). В отличие от Frantext и DMF этот проект долгое время развивается «на общественных началах», благодаря участию аспирантов К. Маркелло-Низья, обменом текстами с коллегами, небольшим грантам, и пополнение корпуса шло медленнее и менее систематично, чем в Нанси. Тем не менее в настоящее время BFM является крупнейшим свободно корпусом старо- и среднефранцузского языка (более 100 текстов и 5 миллионов словоупотреблений), доступ к которому бесплатно предоставляется всем желающим. Бóльшая часть текстов BFM свободно распространяется в форматах XML-TEI [The TEI Consortium] и PDF на условиях лицензии Creative Commons BY-NC-SA (цитирование источника, некоммерческое использование и распространение производных продуктов на тех же условиях, что и исходного). Исключение составляют тексты, имеющие ограничения, связанные с авторскими правами публикаторов и издателей.

Возвращаясь к истории создания исторических корпусов французского языка, мы не можем не упомянуть Амстердамские корпуса грамот и литературных текстов, которые в середине 1980-х годов создает коллектив под руководством профессора А. Дееса для проведения диалектометрических исследований и публикации «Атласа диалектов». Корпус достигал около 3 миллионов словоупотреблений и имел «ручную» морфологическую разметку. После смерти А. Дееса «литературный» корпус долгое время не использовался и едва не был утерян, и лишь в 2006 г. усилиями П. Кунстманна и А. Штайна был переведен в более современный формат, обогащен дополнительной разметкой и опубликован в Интернете. Текстовые данные корпуса грамот удалось сохранить, однако значительная часть разметки, возможно, была утеряна безвозвратно.

2. Проблемы создания и эксплуатации диахронических корпусов

Пример Амстердамского корпуса ярко иллюстрирует опасность потери результатов долгой и кропотливой работы по созданию корпуса при отсутствии плана долгосрочной архивации и постоянной поддержки проекта. В то же время он внушает надежду на то, что заинтересованность научного сообщества и свободное распространение исходных материалов могут способствовать «выживанию» ресурсов после окончания проекта или распада коллектива, их породившего. Следует отметить, что использование международно признанных стандартов кодирования и аннотации текстов существенно повышает шансы на сохранность данных.

Помимо обеспечения долговечности создание и использование корпусов требует решения целого ряда методологических и технических вопросов. Одной из главных проблем корпусной лингвистики является репрезентативность корпусных данных. Как правило, репрезентативность достигается за счет наиболее полного и сбалансированного охвата всех релевантных для конкретного исследования типов продуктов речевой деятельности (текстов и/или записей устной речи). Когда речь идет об исторически удаленных этапах развития языка, набор доступных источников ограничен. Так, всего три небольших текста дошли до нас от французского языка IX – X вв. В подобной ситуации единственным решением является использование всех сохранившихся источников с учетом типологии текстов и статистических методов, позволяющих оценить достоверность обобщений на материале данного корпуса.

При работе с корпусами средневековых текстов необходимо учитывать адекватность представления лингвистических данных в оцифрованных источниках. Следует различать корпуса, построенные из специально подготовленных транскрипций

рукописей, от корпусов, основанных на оцифрованных научных изданиях. Оцифровка научных изданий обходится значительно дешевле и занимает значительно меньше времени, чем качественная транскрипция рукописи, поэтому все «крупные» корпуса средневековых французских текстов (такие, как BFM и база текстов DMF) созданы именно этим методом. Приоритет при выборе источника отдавался тем изданиям, которые наиболее «уважительно» относились к чтениям «базовой» рукописи, однако не все публикаторы строго придерживаются ими же самими заявленных принципов, а часть первичных данных (таких, как пунктуация и графическое словоделение) в традиционных научных изданиях игнорируется полностью. Если для исследований в области лексики или синтаксиса (не говоря уже об истории или литературоведении), эти особенности научных изданий не имеют существенного значения, то для изучения морфологии, графики и пунктуации необходимы более точные, «дипломатические», транскрипции первоисточников. Информационные технологии позволяют создавать синоптические издания, включающие два и более уровня транскрипции, причем со стороны публикатора для их подготовки требуются лишь небольшое дополнительное усилие [La « philologie numérique »...]. Можно надеяться, что со временем подобные издания станут нормой и постепенно заменят традиционные.

Использование традиционных опубликованных изданий при создании корпусов также сопряжено с рядом юридических проблем. Во Франции сложилась практика, согласно которой научные редакторы уступают все права коммерческим издательствам. До 2000 г. в большинстве договоров права на электронные издания в Интернете прямо не оговаривались, поэтому можно считать, что эти права сохраняются за научными редакторами, хотя судебных решений по подобного рода спорам в настоящий момент не существует и вопрос остается спорным. Спорным является и вопрос о применимости авторского права к «базовому тексту» научных изданий средневековой литературы (за исключением примечаний и прочих элементов критического аппарата), однако и здесь какие-либо судебные прецеденты отсутствуют. Договоры, заключенные после 2000 г., как правило, содержат пункт о передаче издателю исключительных прав на электронное распространение издания, что исключает возможность свободного использования текста в корпусах. При этом просьбы о разрешении на включение того или текста в электронный корпус нередко остаются без ответа со стороны издательств. В этих условиях требуется консолидация научного сообщества с тем, чтобы изменить сложившуюся практику и устранить барьеры на пути свободного распространения данных для научных исследований. Создание в 2011 г. консорциумов по развитию корпусов для различных гуманитарных наук во Франции можно считать первым этапом такой консолидации (http://www.huma-num.fr/service/consortium). Кроме того, необходимо создать альтернативу публикации памятников истории языка в частных издательствах, которая сочетала бы свободный доступ к данным с гарантией филологического и эргономического качества издания. Внести вклад в решение этой задачи призвано создание коллекции электронных изданий под эгидой BFM, запланированное на 2014 год. Все издания этой коллекции будут распространяться на условиях «открытой» лицензии типа Creative Commons (http://creativecommons.org). В настоящее время уже опубликован прототип издания «Поисков святого Грааля» [La Queste...], ряд других изданий находится в стадии подготовки.

3. Примеры корпусных исследований

Несмотря на вышеперечисленные проблемы, использование корпусных данных в исследовании истории французского языка активно развивается. Только на материале корпуса BFM защищено не менее 30 диссертаций. Содержание докладов на международных конференциях по истории французского языка (таких, как Diachro и

SIDF) показывает, что анализ корпусов становится неотъемлемой частью большинства исследований.

С 2007 г. по инициативе К. Маркелло-Низья, Б. Комбетта и С. Прево началась работа над созданием «Большой исторической грамматики французского языка» (GGHF), которая призвана прийти на смену классической, но во многом устаревшей 13-томной «Истории французского языка» Ф. Брюно [Brunot]. В настоящее время в проекте участвуют более 10 ведущих специалистов по различным периодам истории и аспектам системы французского языка. Для проекта GGHF был создан специальный корпус, охватывающий всю историю французского языка (IX – XX вв.). В рамках этого корпуса выделен «ядерный» подкорпус, в котором каждый век (начиная с XII) представлен 200 — 250 тысячами словоупотреблений (общий объем — около 2 млн словоупотреблений). Этот подкорпус используется для подсчета и углубленного анализа наблюдаемых явлений. Дополнительный корпус объемом 11 млн словоупотреблений (неравномерно распределенных по векам) используется для поиска примеров и проверки ранее сформулированных гипотез. Электронные тексты для корпуса GGHF были собраны из различных источников (BFM, Frantext, Виртуальная библиотека гуманистов BVH и др.). Созданию единого корпуса способствовало то, что все тексты были размечены в формате XML на основе рекомендаций международного консорциума TEI (http://www.tei-c.org). Эксплуатация корпуса GGHF осуществляется с помощью платформы TXM (http://sourceforge.net/projects/txm), которая разрабатывается в исследовательской лаборатории ICAR и используется в частности Базой средневекового французского языка. Разумеется, корпус GGHF не может считаться репрезентативным согласно определению Дж. Синклера [Sinclair] в силу своих скромных размеров и ограниченного числа представленных текстовых жанров, однако его можно рассматривать как первый этап на пути создания диахронической составляющей национального корпуса французского языка. Как известно, французский язык до сих пор не имеет своего национального корпуса, и лишь в 2012 г. по инициативе Института французской лингвистики (ILF) прошли первые научно-методологические консультации по его составлению.

В качестве примера актуальных корпусных исследований в области французской диахронии приведем несколько работ на материале Базы средневекового французского языка, в которых мы принимали непосредственное участие.

Одним из вопросов, вызывающих в последние годы большой интерес исследователей, является возможность изучения устной речи на материале средневековых рукописей. Известно, что устная традиция играла огромную роль в средние века, когда письмом владела лишь малая часть носителей языка и когда литературные произведения создавались прежде всего для устного исполнения. Можно ли как-то отделить в средневековых текстах феномены, характерные для устной речи, от постепенно формирующихся специфических норм письменного языка? Гипотеза, которую мы хотели исследовать, состоит в том, что прямая речь персонажей литературных произведений, не будучи прямым отражением устной речи, обладает рядом особенностей, существенных с точки зрения лингвистического анализа. Так, по мнению К. Маркелло-Низья, изменения языковой системы проявляются в прямой речи персонажей раньше, чем в «словах автора».

С целью сопоставительного изучения прямой речи и «слов автора» в корпусе старо- и среднефранцузских текстов мы прежде всего разметили с помощью TEI-тэга <q> выделенные в изданиях кавычками участки текста. При всем своем несовершенстве этот метод позволяет быстро получить размеченный корпус солидных размеров с относительно невысоким уровнем «шума». Затем мы использовали метод расчета специфичности и факторного анализа частотности употребления частей речи в прямой

речи и в «словах автора», скрещивая вариацию по этому с другими факторами вариации (диахронической, функционально-стилевой) и индивидуальными особенностями текстов. Результаты исследования, представленные на конференции DIA [L'oral représenté...], показали, что вариация по признаку «прямая речь / слова автора» «перевешивает» все остальные рассмотренные факторы и четко проявляется при любой конфигурации корпуса, что подтверждает первоначально выдвинутую гипотезу.

Другое корпусное исследование, на котором хотелось бы остановиться, связано с понятием предложения в старофранцузском языке. Как мы уже отмечали, в научных изданиях принято использовать современную пунктуацию, с помощью которой научный редактор выражает свой анализ структуры текста. В то же время ряд ученых полагает, что предложений в современном понимании в старофранцузском языке не было, а исследование рукописной пунктуации показывает в первую очередь высокий уровень вариации и невозможность однозначного определения основной «пунктуационной единицы». Тем не менее факторный анализ употребления знаков препинания в рукописях на синтаксических границах, определенных по независимым от графического кода признакам (границы предикативных единиц, наличие подчинительной связи, общих членов, границы прямой речи) показывает, что границы между предикативными единицами, не имеющими общих членов и подчинительной связи, маркируются существенно чаще, чем границы любого другого типа. Это дает основание утверждать, что единица близкая к предложению в современном понимании существовала в старофранцузском языке и находила опосредованное выражение в разнообразной пунктуационной практике той эпохи [Lavrentiev].

Еще один пример корпусного исследования связан с уточнением гипотезы о доминировании в старофранцузском языке порядка слов V2 (при котором глагол-сказуемое занимает вторую позицию в структуре предложения). На материале корпуса синтаксической разметкой мы убедились, что порядок слов V2 действительно доминирует во всех текстах, однако примеры , в которых глагол занимает не вторую, а третью или еще более отдаленную от начала предложения позицию, встречаются везде. Существенное различие наблюдается между прозой и стихотворными текстами, в которых разнообразие конструкций, отступающих от правила V2, значительно шире. В прозе основную массу «исключений» составляет несколько относительно устойчивых конструкций типа «придаточное предложение + *si* ('так') + глагол». Подробнее с результатами этого исследования можно познакомиться в материалах 3-го Мирового конгресса французской лингвистики [La zone préverbale...].

4. Заключение

Приведенные нами примеры показывают, что анализ корпусных данных позволяет получить либо принципиально новые результаты, либо количественно подтвердить и уточнить ранее выдвинутые гипотезы. Открытый доступ к корпусам текстов различных периодов истории французского языка, а в перспективе – создание национального корпуса с диахронической составляющей должны способствовать развитию и качественному росту лингвистических исследований. Обучение методам работы с электронными ресурсами, основам цифровой филологии и правовым аспектам издания памятников письменности должны стать неотъемлемой частью филологического образования.

Литература

Brunot F. Histoire de la langue française des origines à nos jours. Paris : A. Colin, 1905–1953. 13 vol.

Foulet L. Petite syntaxe de l'ancien français. Paris : Champion, 1919.

L'oral représenté dans un corpus de français médiéval (9e–15e) : approche contrastive et outillée de la variation diasystémique : [докл. межд. конф. «ΔIA II : Les variations diasystémiques et leurs interdépendances» (Копенгаген, 19-21 ноября 2012 г.)] / C. Guillot [и др.]. // HAL-SHS [электронный архив]. URL: http://halshs.archives-ouvertes.fr/halshs-00760647 (дата обращения 17.11.2013).

La "philologie numérique": tentative de définition d'un nouvel objet éditorial : [докл. межд. конф. «27e Congrès international de philologie et de linguistique romanes» (Нанси, Франция, 15-20 июля 2013 г.)] / C. Guillot [и др.]. // HAL-SHS [электронный архив]. URL: http://halshs.archives-ouvertes.fr/halshs-00846767 (дата обращения 17.11.2013).

La Queste del saint Graal : Édition numérique interactive du manuscrit Lyon, BM, P.A. 77 / éd. par C. Marchello-Nizia et A. Lavrentiev. Lyon : ENS de Lyon, 2013. Доступ через портал «BFM» : http://txm.bfm-corpus.org (дата обращения 17.11.2013).

La zone préverbale en ancien français : apport des corpus annotés / T. Rainsford [и др.] // CMLF 2012 : 3ème Congrès Mondial de Linguistique Française (Lyon, France, 2012). DOI: 10.1051/shsconf/20120100246. URL: http://dx.doi.org/10.1051/shsconf/20120100246 (дата обращения 17.11.2013).

Lavrentiev A. La 'phrase' en français médiéval : une réalité ou une reconstruction artificielle ? // Actes du CMLF 2010 : 2ème Congrès Mondial de Linguistique Française (La Nouvelle Orléans, États-Unis, 2010). DOI: 10.1051/cmlf/2010125. URL: http://dx.doi.org/10.1051/cmlf/2010125 (дата обращения 17.11.2013).

Lecoultre J.-J. De l'Ordre des mots dans Crestien de Troyes. Dresden : B. G. Teubner, 1875.

Marchello-Nizia C. Histoire de la langue française aux XIVe et XVe siècles. Paris: Bordas, 1979.

Sinclair J. Preliminary Recommendations on Corpus Typology. EAGLES, May 1996. URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (дата обращения 17.11.2013).

The TEI Consortium. TEI P5 : Guidelines for Electronic Text Encoding and Interchange / ed. by L. Burnard and S. Bauman. Charlottesville, Virginia: TEI, 2013. URL : http://www.tei-c.org/Guidelines/P5/ (дата обращения 16.11.2013).