Текстометрический инструментарий в исследовании средневековой пунктуации

Alexei Lavrentiev

УДК 801.82:004.9

А.М. Лаврентьев, кандидат филологических наук, доктор Лионского университета, научный сотрудник лаборатории ICAR, Национальный центр научных исследований Франции (CNRS)

ТЕКСТОМЕТРИЧЕСКИЙ ИНСТРУМЕНТАРИЙ В ИССЛЕДОВАНИИ СРЕДНЕВЕКОВОЙ ПУНКТУАЦИИ

Рассмотрены возможности использования поисково-аналитической машины Weblex для проведения лингвистического исследования на материале корпуса транскрипций средневековых рукописей, содержащих аналитическую и лингвистическую разметку на основе стандарта XML TEI <http://www.tei-c.org>. Объектом исследования была эволюция пунктуации в рукописях французских прозаических текстов XIII–XV веков.

Ключевые слова: *текстометрия, корпусная лингвистика, транскрипция рукописи, TEI*

Использованный в настоящем исследовании корпус включает 28 «многоуровневых транскрипций» фрагментов рукописей объемом от 550 до 2250 текстоформ. Общий объем корпуса составляет около 28 100 текстоформ. XIII век представлен шестью рукописями, а XIV век – пятью. XV веком датируются 13 рукописей и 2 инкунабулы. Кроме того, в корпус включены две печатные книги первой половины XVI в.

Все тексты корпуса детально описаны в соответствии с принятой в Базе средневекового французского <http://bfm.ens-lsh.fr> системой типологической классификации. Важнейшим параметром типологического описания текста является его принадлежность к определенной функциональной сфере. В использованном корпусе по девять текстов относятся к литературной и к научно-дидактической сферам, шесть текстов – к исторической и два – к религиозной. По одному тексту представляют юридическую и политическую сферу.

Разумеется, данный корпус никоим образом не может считаться репрезентативным, однако он позволяет проследить наиболее общие тенденции и сформулировать гипотезы для дальнейшего исследования на более обширном материале.

Исследование пунктуации, так же как и прочих элементов графической системы рукописей, требует от транскрипции точного представления данных первоисточника. Вместе с тем определенное упрощение и нормализация данных необходимы для удобства чтения и автоматической обработки текста (например, для лемматизации и морфологической разметки).

Решить подобную задачу позволяет многоуровневая транскрипция с синхронизацией различных уровней представления данных на уровне словоформы. Принципы подобной транскрипции были впервые обоснованы и реализованы в рамках проекта «Архива средневековых нордических текстов» (Menota) <http://www.menota.org> [Haugen 2004].

Действующие в настоящее время рекомендации XML-TEI (P5) содержат подавляющее большинство элементов, необходимых для кодировки полноценных многоуровневых транскрипций средневековых европейских рукописей и их описаний. Эти элементы содержатся в модулях «общего

назначения»: *TEI, analysis, core, corpus, header, linking, namesdates* и *textstructure*, а также в «специализированных» модулях: *gaiji, msdescription* и *transcr*. Подробное описание модулей и входящих в них элементов представлено на сайте <http://www.tei-c.org> в разделе «Guidelines».

Список тэгов, использованных нами при создании корпуса для изучения пунктуации, представлен в «Руководстве по кодированию рукописей» проекта «BFM-рукописи» [Lavrentiev 2008].

Вместе с тем специфика нашего проекта обусловила необходимость введения нескольких дополнительных элементов. В первую очередь, речь идет об элементах, позволяющих идентифицировать три уровня транскрипции словоформы (или знака препинания). Соответствующие элементы – <me:norm>, <me:dipl> и <me:facs> – заимствованы нами из схемы, разработанной в проекте *Menota*. В структуре XML-разметки они помещаются внутрь элемента *choice* (указывающего на то, что речь идет об альтернативных формах представления одного и того же сегмента текста), который, в свою очередь, помещается в элемент <w> (слово) или <bfm:punct> (знак препинания).

Корпус снабжен аналитической разметкой нескольких видов. Ключевой аннотируемой единицей является «пунктуационная граница». Под пунктуационной границей понимается стык составляющих текст синтагматических единиц, на котором вероятно появление пунктуации. На первоначальном этапе состав пунктуационных границ определяется эмпирически: проводится анализ условий, в которых знаки препинания встречаются в различных текстах, а затем аналогичные позиции выявляются и размечаются в текстах систематически, независимо от присутствия знаков препинания. По мере накопления фактического материала перечень пунктуационных границ стабилизируется, и лишь очень незначительная доля употреблений остается «за рамками» выделенных пунктуационных границ. В этих случаях нельзя исключить действия экстралингвистических факторов (таких, как выравнивание конца строки) или ошибки писца.

Аннотация пунктуационной границы состоит в указании «формы» и «силы» знака препинания, а также типа границы. При идентификации формы знака препинания различаются графемы и аллографы. В нашем корпусе к числу графем относятся точка (независимо от ее положения на строке), косая черта, вопросительный знак, комма и несколько других знаков. В качестве примера аллографов можно привести «длинный» и короткий вариант косой черты или «прямую» и «закругленную» форму коммы. Особый код используется для обозначения отсутствия выраженного отдельным графическим сегментом знака препинания. Заметим, что употребление на пунктуационной границе прописной буквы рассматривается нами как форма пунктуации и при отсутствии знака препинания как такового.

«Сила» пунктуации определяется оформлением следующего за пунктуационной границей сегмента текста. В случае, когда сегмент начинается с новой строки (при незаполненной предыдущей) и/или с буквицы, можно говорить об «экстрасильной» пунктуации. Сильная пунктуация определяется употреблением прописной буквы. При этом наличие или отсутствие знака

препинания роли не играет. Слабая пунктуация возникает при употреблении какого-либо знака препинания с последующей строчной буквой.

Тип пунктуационной границы определяется рядом факторов: иерархическим уровнем стыкующихся единиц, их тематической близостью, принадлежностью к авторскому дискурсу или к цитатам. С точки зрения синтаксической иерархии границы можно разделить на «горизонтальные» и «вертикальные». В первом случае речь идет о стыке единиц одного уровня (например, простых предложений с сочинительной или бессоюзной связью или однородных членов предложения), во втором мы имеем дело с выделением зависимых единиц в составе более крупных (например, подчиненное предложение в составе сложного или обособленный член предложения). Вертикальные границы могут быть «левыми» или «правыми» (в начале или в конце зависимой единицы соответственно).

По завершении разметки корпус загружается в поисково-аналитическую машину Weblex <http://weblex.ens-lsh.fr/wlx> с целью его дальнейшего исследования. Процедура загрузки корпуса в Weblex является достаточно сложной и может осуществляться только администратором в пределах локальной сети. В рамках проекта Текстометрия <http://texmotetrie.ens-lsh.fr>, финансируемого французским Национальным агентством исследований (ANR), разрабатывается платформа TXM, призванная прийти на смену Weblex. Она будет функционировать как в локальном, так и в сетевом режиме, а процедура интеграции текстов, размеченных согласно нормам XML-TEI, существенно упростится. Платформа TXM разрабатывается на основе открытого кода (open source). Ее наиболее актуальная версия может быть загружена с сайта <http://sourceforge.net/projects/textometrie/>.

Среди множества предоставляемых Weblex функциональных возможностей «традиционными» лингвистами особенно востребованы следующие:

1. **Вокабуляр**. Эта функция позволяет получать алфавитный и частотный список словоформ корпуса или их атрибутов. В качестве атрибутов могут выступать лемма или какая-либо из форм разметки. В нашем корпусе атрибут P2 использовался для аннотации типа текстоформы: словоформа или пунктуация (с указанием силы). Использование функции «вокабуляр» по данному атрибуту позволяет легко вычислить относительную частотность сильной и слабой пунктуации в отдельном тексте.

2. **Индекс форм**, отвечающих запросу. Weblex и платформа TXM поддерживают язык запросов CQP, разработанный в штуттгартском Институте машинной лингвистики <http://cwb.sourceforge.net/>. Синтаксис CQP позволяет формулировать достаточно сложные, комбинирующие различные атрибуты отдельной текстоформы или последовательности текстоформ. Даже при работе с нелемматизированным корпусом средневековых текстов с их высокой вариативностью словоформ можно добиться достаточно хорошего соотношения «шума» и «тишины» в результатах запросов. В нашей работе функция индекса использовалась в частности для получения частотного списка сочетания типа пунктуационной границы, формы и силы пунктуации.

3. **Конкорданс KWIC**. Данная функция позволяет получить конкорданс текстоформ, отвечающих запросу, с возможностью сортировки по различным полям (форма, левый или правый контекст, ссылка). В отличие от многих других программ Weblex не ограничивает размеры контекста (хотя при запросе контекстов в несколько тысяч текстоформ могут возникнуть проблемы, связанные с нехваткой аппаратных ресурсов). Конкордансы необходимы для проведения «тонкого» анализа отдельных текстоформ.

Среди чисто количественных методов, позволяющих получить интересные результаты при анализе корпуса в целом, необходимо упомянуть **факторный анализ** [Харман 1972]. Факторный анализ позволяет сократить число представленных в статистических данных переменных и выявить взаимосвязи между ними. Результаты такого анализа часто представляются в виде двухмерного графика, или факторного плана. В нашем случае мы проанализировали данные по сочетанию силы пунктуации с определенным типом границы на «популяции» текстов корпуса. При этом полученные с помощью Weblex данные были обработаны с помощью программы R. Заметим, что в новой платформе TXM функция факторного анализа интегрирована.

Факторный анализ наглядно демонстрирует, что пунктуационное оформление границ автономных предложений (C1) четко противопоставляется остальным рассмотренным типам границ (предложения с общими членами, придаточные предложения, однородные члены). Также более или менее отчетливо формируются три группировки текстов. Одна из них характеризуется общим низким уровнем пунктуации, другая – преобладанием сильной пунктуации, а третья – сравнительно высоким уровнем пунктуации с некоторым преобладанием слабой. Данные выводы подтверждаются тщательным качественным анализом употребления пунктуации. Два текста явно «выпадают» из общей массы. Это трактат Этьена Доле (Étienne Dolet) «О правильной манере перевода» (La manière de bien traduire d'une langue en aultre, 1540) и «Страсбургские клятвы» (Les serments de Strasbourg /Sacramenta Argentariae/) от 14 февраля 842 года. Эти тексты действительно занимают в корпусе особое положение как по «внешним признакам» (прежде всего по дате), так и по тенденциям пунктуации.

В целом можно сделать вывод, что факторный анализ дает на нашем материале достаточно корректные результаты и может быть успешно применен на более обширном и репрезентативном корпусе.

В качестве заключения приведем основные результаты исследования тенденций пунктуации на материале нашего корпуса.

Первым параметром, учитываемым при характеристике пунктуации источника, является ее «общий уровень», измеряемый в среднем количестве знаков препинания на 100 слов текста. Для удобства мы обозначаем этот уровень в процентах. В большинстве рукописей XIII и XIV вв. уровень пунктуации варьируется от 8 до 10 процентов. Встречаются и исключения: в двух рукописях общий уровень пунктуации достигает 13 процентов, а в одной

рукописи, напротив, не превышает 3 процентов. Заметим, что в современных французских текстах этот уровень достигает 14–16 процентов.

В XV в., по всей видимости, происходит незначительное снижение среднего уровня пунктуации. В большинстве источников он колеблется от 4 до 8 процентов, при этом в одной рукописи он опускается до 3 процентов, а в другой поднимается до 10 процентов.

Анализ формы знаков препинания показывает, что, как правило, в тексте доминирует один знак препинания (точка в 15 рукописях и одной инкунабуле, косая черта в двух рукописях и одной книге XVI в.), но всегда эпизодически встречается как минимум один другой. В четырех рукописях наряду с отдельно стоящими знаками препинания широко используется прописная буква без знака препинания. Эта практика особенно распространена в XV в. Наконец, в двух рукописях и в двух печатных книгах несколько различных знаков препинания используется с сопоставимой частотностью. При этом, однако, один из знаков все же преобладает.

Характерной чертой средневековой пунктуации является отсутствие четкой «специализации» знаков. Один и тот же знак может использоваться как в сильной, так и в слабой пунктуации и не связан с каким-либо определенным типом пунктуационной границы. Лишь в некоторых рукописях появляются знаки, ориентированные на определенную функцию. Так, вопросительный знак может указывать на вопросительную модальность или на эмоциональную маркированность высказывания.

Чаще всего пунктуация встречается на границе автономных предикативных единиц. Заметно реже знаки препинания используются на границах предикативных единиц с общими компонентами и на границах придаточных предложений. Однородные члены предложения разделяются пунктуацией, как правило, в длинных перечислениях, особенно если перечисляются имена собственные.

Если в тексте присутствует прямая речь, то достаточно регулярно пунктуация присутствует в ее начале и в конце, а также при смене реплики в диалоге. Напротив, вводные «слова автора» крайне редко выделяются пунктуационно. Обособленные синтагмы оформляются пунктуацией редко и лишь в части рукописей.

В целом можно сделать вывод о том, что, несмотря на кажущуюся нерегулярность их употребления, знаки препинания появляются во французских средневековых текстах во вполне определенных синтаксических позициях с учетом ряда семантических и экстралингвистических факторов.

Список литературы

1. *Копотев М. В., Мустайоки А.* Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет (Principles of the Creation of the Helsinki Annotated Corpus HANCO) // Научно-техническая информация. Сер. 2, Информационные системы и процессы. – 2003. – № 6. Корпусная лингвистика в России. – С. 33–37.

2. *Харман Г.* Современный факторный анализ / пер. с англ. В. Я. Лумельского. – М. : Статистика, 1972. – 486 с.

3. *Haugen, O. E.* Parallel Views: Multi-level Encoding of Medieval Nordic Primary Sources // Literary and Linguistic Computing. – 2004. – Vol. 19, nr 1. – Pp. 73-91.

4. *Lavrentiev, A.* Manuel d'encodage XML-TEI étendu des transcriptions de manuscrits dans le projet BFM-Manuscrits, v. 2.1. Lyon : UMR ICAR, 2008. – Адрес в Интернет http://ccfm.ens-lsh.fr/IMG/pdf/BFM-Mss_Encodage-XML.pdf.

A. Lavrentiev

Textometric tools for a research on medieval punctuation

This paper describes the way how Weblex search and analysis engine can be used for a linguistic research on a corpus medieval manuscript transcriptions endoded according to XML-TEI standard. The research focused on the evolution of punctuation in French prose texts from the 13th to the 15th centuries.

Keywords: *textometry, corpus linguistics, punctuation, manuscript transcription, TEI*

A. Lavrentiev

Ученая степень, ученое звание, должность, место работы: Ph. D. in Linguistics, Research engineer, ICAR research laboratory, CNRS, Lyon, France

CNRS – the *Centre National de la Recherche Scientifique* (National Center for Scientific Research)

Название статьи: Textometric tools for a research on medieval punctuation