Лаврентьев А.М.
Высшая нормальная школа гуманитарных наук /
Лионский университет

Исследование пунктуации и графической сегментации средневековых рукописей с использованием электронных транскрипций и поисковой машины Weblex

Чем интересна средневековая пунктуация

Долгое время пунктуация и словоделение западноевропейских рукописей не привлекали большого внимания исследователей. Считалось, что пунктуация в этих рукописях либо отсутствовала, либо использовалась бессистемно.

В научных изданиях средневековых текстов пунктуации первоисточника уделялось от силы несколько строчек во введении, а в сам текст публикаторы без колебания вводили современную пунктуацию. Подобная практика, безусловно, облегчает чтение и понимание текста (вернее, его интерпретации публикатором), однако неизбежно скрывает от читателя определенную часть данных и возможных интерпретаций первоисточника.

В отсутствие нормативной грамматики и общеизвестного свода правил «спонтанная» пунктуация средневековых рукописей отражает «наивные» представления носителей языка о структуре текста и иерархии его компонентов. Эта информация может быть весьма полезной при изучении синтаксиса древнего языка, процессов грамматикализации определенных лексем и конструкций. Практика словоделения несет информацию о синтагматической автономности отдельных морфем и о формировании лексических единиц из свободных словосочетаний.

Хотя ни пунктуация, ни графическая сегментация не отражают напрямую ритмику и интонацию устной речи, они могут служить одним из немногих (если не единственным) источником информации об этих аспектах речевой деятельности в периоды, когда не существовало звукозаписывающей аппаратуры.

Немалый интерес представляет пунктуация и для палеографии. Наряду с почерком и художественным оформлением рукописи пунктуация может помочь в идентификации рукописной традиции, определении места и времени ее написания.

В 1952 г. известный французский романист Марио Рок (Mario Roques) обратил внимание на богатую пунктуацию знаменитого писца Гийо (Guiot), «редактировавшего» в XIII в. романы Кретьена де Труа, и высказал пожелание о проведении более широкого исследования истории французской пунктуации в текстах различных типов [Roques 1952]. Четверть века спустя Кристиана Маркелло-Низья (Christiane Marchello-Nizia) публикует статью [Marchello-Nizia 1978], содержащую первое глубокое лингвистическое исследование практики средневековой французской пунктуации. В статье Маркелло-Низья, построенной на данных анализа 7-й главы романа Жана де Бёля (Jean de Bueil) «Le Jouvencel» в редакциях пяти рукописей XV в. и одной инкунабулы, сформулирован и рассмотрен ряд гипотез относительно

лингвистических структур, скрывающихся за фрагментами текста, отмеченными пунктуацией, или «единицами чтения» (unités de lecture), согласно терминологии исследовательницы.

Несмотря на ряд полученных ценных результатов, статья заканчивается достаточно пессимистическим выводом о невозможности их обобщения ввиду «индивидуальности» каждой рукописи и о «недостаточности» современного инструментария лингвистического анализа для адекватного описания «единиц чтения».

Тридцать лет спустя после выхода в свет труда Маркелло-Низья на тему средневековой французской пунктуации и словоделения опубликовано несколько статей [Buridant 1980; Barbance 1995; Andrieux-Reix и Monsonégo 1997; Baddeley 2001 и др.] и защищены четыре диссертации, в той или иной мере посвященные этим проблемам [Llamas Pombo 1996; Li 2007; Mazziotta 2007; Lavrentiev 2009], однако многие вопросы по-прежнему остаются открытыми. Так, до сих пор не удалось определить, была ли в развитии практики пунктуации одна «генеральная линия», от которой отдельные писцы могли достаточно существенно отклоняться, или же существовало несколько параллельных независимых традиций.

Проблемы в изучении пунктуации

Сложность изучения средневековой пунктуации связана прежде всего с необходимостью накопления достаточного объема эмпирических данных. Крупные корпусы средневековых текстов (такие, как «База средневекового французского»[1] или «Старофранцузские тексты»[2]) созданы на основе научных изданий с нормализованной пунктуацией, транскрипция же и анализ первоисточников требуют больших затрат времени и материальных вложений.

Необходимость привлечения достаточно крупного и диверсифицированного корпуса для изучения средневековой пунктуации связана с ее высокой вариативностью, отмеченной всеми ее исследователями. В этих условиях всякое обобщение результатов, основанных на данных ограниченного числа источников, представляется малоубедительным. Между тем, подавляющее большинство опубликованных до настоящего времени исследований основано на материале от одного до пяти источников.

Дополнительная сложность в исследовании пунктуации связана с малым размером знаков препинания. При работе с фотографиями рукописей, а иногда и с оригиналами бывает сложно отличить знак препинания от пятна или дефекта поверхности. Кроме того, иногда практически невозможно определить, кем поставлен знак препинания: писцом, его современником, корректировавшим рукопись, или же каким-либо читателем в более позднее время. Наконец, идентификация того или иного знака препинания по его форме может в отдельных случаях вызывать затруднения. Так, восходящие к разным рукописным традициям знаки comma и punctus elevatus по форме практически неотличимы друг от друга [Parkes 1992: 303, 306], а порой «смешиваются» и с punctus interrogativus. То же можно сказать и о знаках абзаца crochet adlinéaire и pied-de-mouche, восходящих соответственно к греческой букве Γ и к латинской C [Parkes 1992: 305]. Следует, впрочем, отметить, что форма знака имеет в целом меньшее значение для лингвистического анализа, чем его наличие или отсутствие, а также чем форма последующей буквы (строчная или прописная).

[1] Base de Français Médiéval (BFM), <http://bfm.ens-lsh.fr>.
[2] Textes de Français Ancien (TFA), <http://www.lib.uchicago.edu/efts/ARTFL/projects/TLA/>.

Несмотря на все сложности, средневековая пунктуация заслуживает изучения, и тайны ее будут открываться по мере накопления эмпирических данных и совершенствования методики их анализа.

В наши дни в области научного издания старых текстов происходит настоящая революция, связанная с внедрением информационных технологий. В электронном издании форма представления текста источника и сложность научного аппарата могут адаптироваться к запросам читателя или исследователя, тем самым снимается давняя дилемма, состоявшая в выборе между точным отражением данных источника и удобством для современного читателя. Вместе с тем стабильная научная методика и технологические стандарты, призванные обеспечить качество и долговечность подобных изданий, еще не сложились, и существует риск, что дорогостоящее электронное издание или корпус текстов станет через несколько лет после его создания практически непригодным для использования.

Шансы на долговечность электронного издания могут существенно повыситься при условии, что оно учитывает интересы различных групп потенциальных пользователей, основано на открытых технологиях и опирается на широко распространенные в области кодирования и разметки текстов международные стандарты. В настоящее время подобным требованиям в наибольшей мере отвечает язык XML и схема разметки, разработанная в рамках консорциума Text Encoding Initiative (TEI). В дальнейшем мы будем использовать аббревиатуру XML-TEI для обозначения данной схемы разметки.

В рамках посвященного тенденциям пунктуации во французских рукописях в прозе XIII–XV вв. диссертационного исследования мы сформировали корпус транскрипций фрагментов рукописей и инкунабул соответствующего периода, основанных на принципе трехуровневого представления данных, в расширенном формате XML-TEI.

Уровни транскрипции

Для иллюстрации принципа трехуровневой транскрипции средневековых рукописей и инкунабул воспользуемся примером «Страсбургских клятв», традиционно считающихся самым древним памятником французского языка, дошедшим до наших дней. Этот небольшой текст представляет собой клятвы, произнесенные в 842 г. внуками Карла Великого, Карлом Лысым и Людовиком Немецким, перед своими дружинами на понятном им языке (старофранцузском и древневерхненемецком). Клятвы эти цитируются дословно в латиноязычной хронике Нитгарда, сохранившейся в рукописи, датирующейся рубежом X–XI веков.

Фотография текста клятвы, произнесенной Людовиком Немецким, приводится на Рис. 1.

Ввиду колоссального символического значения франкоязычной части «Клятв» в истории французского языка, ее текст неоднократно переиздавался в различных хрестоматиях и исследовательских работах, причем особое внимание уделялось точности отражения источника в транскрипции. Так, в хрестоматии «Textes d'étude», составленной Л. Вагнером в 1949 г., а затем исправленной и дополненной О. Колле в 1995 г. [Textes… 1995], представлено две транскрипции «Клятв», «дипломатическая» и «традиционная». В дипломатической транскрипции воспроизведены символы сокращений, словоделение и пунктуация источника, тогда как в традиционной транскрипции сокращения развернуты, словоделение и пунктуация нормализованы. В то же время, вопреки сложившейся практике, не вводится различение букв *u* и *v* по

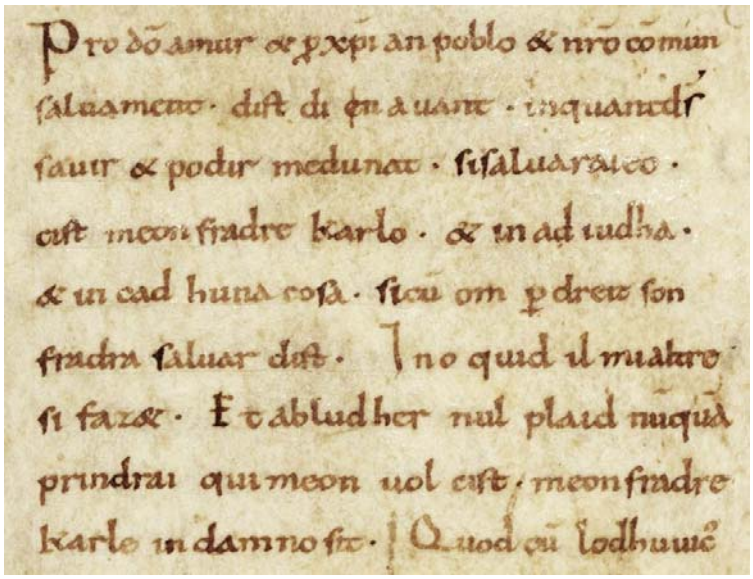

Рис. 1. Фрагмент «Страсбургских клятв» (рукопись, Париж, BnF, lat. 9768). Фотография BnF.

признаку «гласный» / «согласный». В прочих изданиях варьируется тот или иной аспект транскрипции, приближаясь либо к воспроизведению формы источника, либо к современным издательским нормам. При этом публикаторы, как правило, не утруждают себя объяснением причин своего выбора.

Между тем, по сути различные варианты транскрипции связаны с определенными уровнями лингвистического анализа и интерпретации. Если при записи устной речи принято говорить об орфографической, фонематической и фонетической транскрипции, в транскрипции рукописи можно аналогичным образом выделить нормализованный, графематический и аллографический уровни.

Если при транскрипции средневековых текстов речь, естественно, не может идти о приведении написания слов к единой орфографической норме, сложившаяся практика научных изданий предполагает целый ряд операций по нормализации текста. Помимо уже упомянутых модернизации пунктуации и словоделения, развертывания сокращений и введения оппозиций *u / v* и *i / j* практикуется нейтрализация вариантов букв (таких, как «длинное *s*» или «круглое *r*») и введение ряда диакритических знаков (например, «острого акцента» над ударным *e* в конце слова: *parlé*). Также нормализуется использование прописных букв: они употребляются в начале предложений и имен собственных. Образец нормализованной транскрипции представлен на Рис. 2.

Pro Deo amur et pro christian poblo et nostro commun
salvament, d'ist di in avant, in quant Deus
savir et podir me dunat, si salvarai eo
cist meon fradre Karlo et in aiudha
et in cadhuna cosa, si cum om per dreit son
fradra salvar dift, in o quid il mi altresi
fazet. Et ab Ludher nul plaid nunqua·m
prindrai qui, meon vol, cist meon fradre
Karle in damno sit.

Рис. 2. «Нормализованная» транскрипция «Страсбургских клятв»

Подобная нормализация облегчает чтение и анализ текста человеком, а также способствует его более эффективной машинной обработке (в частности, автоматической разметке слов и предложений, лемматизации и морфологической разметке). В то же время нормализация приводит к потере информации об исходной пунктуации и «маскирует» часть свойственной рукописному тексту естественной неоднозначности. Так, при развертывании сокращений публикаторы могут по своему усмотрению восстанавливать падежно-числовой показатель *-s*. Использование данных таких изданий при изучении эволюции именного склонения может, таким образом, привести к получению некорректных или необоснованных результатов.

При необходимости более точно отразить данные источника и в то же время не «перегружать» транскрипцию слишком детальной информацией может использоваться «графематический» уровень. Принцип графематической транскрипции – отражение только существенных с точки зрения языковой системы графических оппозиций. На этом уровне «нейтрализуются» позиционные и каллиграфические варианты букв. Ввиду того, что латинский алфавит не претерпел серьезных изменений со времен средневековья до наших дней, фактически речь идет об использовании букв современного алфавита. Сокращения могут быть развернуты, но «восстановленные» буквы должны графически отличаться от букв источника (например, с помощью выделения курсивом). Диакритические знаки и дополнительные оппозиции букв не применяются. Прописные и строчные буквы употребляются, как в источнике; точка и запятая используются для передачи сильной и слабой пунктуации рукописи или инкунабулы соответственно. Дополнительные знаки (например, вопросительный) могут использоваться, если они употребляются в источнике с четко выраженной функцией. Словоделение на этом уровне транскрипции нормализуется, однако желательно также отражать информацию о реальной графической сегментации источника. В нашей работе мы используем знак «+» при графический агглютинации (слитном написании нескольких слов) и знак «_» при деглютинации (употреблении пробела внутри слова). Агглютинация, связанная с фонетической элизией, обозначается апострофом, как и в нормализованной транскрипции. Графематическая транскрипция «Страсбургских клятв» приведена на Рис. 3.

Pro d*e*o+amur *et*+p*ro*+*christ*i_an poblo *et* n*ost*ro co*m*mun
saluament, d'ist di In auant, in+quant+d*eu*s
sauir *et* podir me+dunat, si+saluarai+eo,
cist meon fradre karlo, *et* in aiudha,
et in cad_huna cosa, si+cu*m* om p*er* dreit son
fradra saluar dift. I_n+o quid il mi+altresi
faz*et*. _Et+ab+ludher nul plaid nu*n*qua·*m*
prindrai qui+meon uol cist meon+fradre
karle in+damno+sit.

Рис. 3. «Графематическая» транскрипция «Страсбургских клятв»

Графематическая транскрипция предоставляет надежные данные для большинства видов лингвистических исследований. Ее подготовка не требует существенно больших затрат времени, чем нормализованная транскрипция, и хотелось бы верить, что со временем предоставление графематической транскрипции станет наряду с нормализованной неотъемлемым элементом электронных изданий.

Для исследований, направленных на изучение графических систем рукописей и в частности пунктуации, графематической транскрипции может быть недостаточно. Символы сокращений и различные варианты букв должны быть отражены в транскрипции. Разумеется, транскрипция не может быть простой имитацией наблюдаемых в источнике графических знаков. Транскрипция всегда предполагает интерпретацию и определенный уровень обобщения, который зависит от целей, для которых транскрипция предназначена. Точное описание деталей начертания букв и знаков препинания может представлять ценность для палеографических исследований, однако с лингвистической точки зрения эти детали интереса не представляют. Наиболее глубоким лингвистически значимым уровнем транскрипции является аллографический, на котором учитываются потенциально функциональные варианты графем. Так, например, оппозиция «короткого» и «длинного» *s* (‹s› / ‹ſ›) приобретает в определенный момент четко выраженный позиционный характер (короткое *s* используется на конце слова, а длинное – во всех остальных позициях). Употребление короткого *s* становится, таким образом, дополнительным средством словоделения и в некоторых случаях (особенно в инкунабулах) заменяет пробел. Критерий «потенциальной значимости», позволяющий отличить лингвистически существенные варианты от несущественных, в известной мере субъективен, однако по мере накопления знаний о функционировании графических систем на разных этапах истории объективность принимаемых решений повышается. В настоящее время можно говорить о том, что состав аллографов, встречающихся в средневековых европейских рукописях, определен достаточно хорошо, о чем свидетельствует успех международной инициативы[3] по созданию стандартной кодировки подобного рода символов.

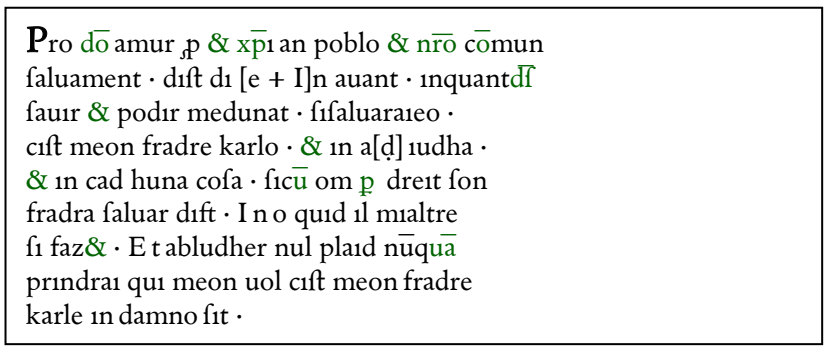
Pro dō amur ꝑ & xp̄ı an poblo & nr̄o c̄omun
ſaluament · dıſt dı [e + I]n auant · ınquantdſ̄
ſauır & podır medunat · fıſaluaraıeo ·
cıſt meon fradre karlo · & ın a[d] ıudha ·
& ın cad huna coſa · ſıc̄u om ꝑ dreıt ſon
fradra ſaluar dıft · I n o quıd ıl mıaltre
ſı faz& · E t abludher nul plaıd nū̄qua
prındraı quı meon uol cıſt meon fradre
karle ın damno ſıt ·

Рис. 4. «Аллографическая» транскрипция «Страсбургских клятв»

Подготовка аллографической транскрипции требует больших затрат времени и может существенно затянуть процесс публикации текстов значительного объема. В то же время следует заметить, что исследования, для которых необходима подобная транскрипция, как правило, не требуют привлечения большой массы текстовых данных одного источника. По этой причине представляется целесообразным ограничить для «больших» текстов объем аллографической транскрипции несколькими фрагментами объемом до 2000 текстоформ[4], тогда как нормализованная и графематическая транскрипции должны быть полными.

[3] Medieval Unicode Font Initiative (MUFI), <http://www.mufi.info>.
[4] Здесь и далее для обозначения словарных единиц корпуса мы будем использовать термин, предложенный М.В. Копотевым и А. Мустайоки [2003].

Кодирование транскрипции

Практическая реализация многоуровневой транскрипции требует решения вопроса о синхронизации. Возможно, разумеется, подготовить три независимых транскрипции всего текста или его крупных фрагментов, однако это существенно осложнит для пользователя переход от одного уровня к другому и сделает невозможным создание «гибридных» форм представления (например, графематическое представление слов и аллографическое представление знаков препинания).

Наибольшую «гибкость» для пользователя предоставляет синхронизация на уровне минимальных различительных единиц графической системы, т.е. графем и диакритических знаков. Технически, однако, подобная синхронизация достаточно сложна, и соответствующая транскрипция практически не будет поддаваться чтению без использования специального программного обеспечения.

Разумным компромиссом представляется синхронизация на уровне словоформы, позволяющая достичь достаточной гибкости формы представления в сочетании с легкостью обработки и «прозрачностью» кодировки. Принципы многоуровневой транскрипции с синхронизацией на уровне слова 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], а также в документе спецификации TEI «bfmmss_d_odd.xml», содержащемся в материалах к данной лекции.

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>[5], <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).

Последний элемент входит в число введенных нами дополнительных по отношению к предлагаемому TEI набору тэгов и служит для разметки знаков препинания. В апреле 2009 г. Совет TEI (TEI Council) принял наше предложение о создании специального элемента для знаков препинания, который получил наименование <pc> (punctuation character). В перспективе этот новый тэг заменит использованный нами <bfm:punct>.

Помимо элемента для пунктуации мы ввели в схему разметки следующие дополнительные элементы:

1) <bfm:lettrine> (для разметки буквиц);

[5] Префикс me: указывает на принадлежность к пространству имен <http://www.menota.org/ns/1.0>; префикс bfm: указывает на принадлежность к пространству имен <http://bfm.ens-lsh.fr/ns/1.0>.

```
<lb n="1"/>
<w xml:id="w1030_001">
    <choice>
        <me:norm>Pro</me:norm>
        <me:dipl>Pro</me:dipl>
        <me:facs>
           <bfm:dropcap size="1" color="black">P</bfm:dropcap>ro
        </m:facs>
    </choice>
</w>
<w xml:id="w1030_002">
    <choice>
        <me:norm>Deo</me:norm>
        <me:dipl>d<ex>e</ex>o</me:dipl>
        <me:facs>
           <bfm:mdvAbbr>
            do&bar;
           </bfm:mdvAbbr>
        </me:facs>
    </choice>
</w>
```

Рис. 5. Фрагмент «расщепленной» кодировки многоуровневой транскрипции

2) <bfm:mdvAbbr> (для разметки средневековых сокращений, этот элемент обладает рядом отличий от стандартного <abbr>);

3) <bfm:headlb> (для разметки «новой строки» в заголовках, «перекрещивающихся» с концом предыдущего раздела);

4) <bfm:hyphen> (для обозначения знака переноса слова на новую строку, присутствующего в источнике);

5) <bfm:sb> (для обозначения деглютинации, или «пробела внутри слова»).

Кроме того, к стандартному набору атрибутов элемента <w> были добавлены @bfm:aggl и @bfm:agglCert. Первый из них позволяет отмечать случаи агглютинации (отсутствия пробела между словами). Второй атрибут используется в тех случаях, когда наличие или отсутствие пробела не является очевидным, в частности, при употреблении в рукописи «малого пробела».

Целесообразность введения дополнительных элементов и атрибутов обоснована в уже упомянутом «Руководстве по кодированию рукописей».

На Рис. 5 показан образец разметки многоуровневой транскрипции двух первых слов «Страсбургских клятв». Подобную систему разметки, в которой каждый из уровней транскрипции представлен отдельным элементом XML, можно назвать «расщепленной». Она легко преобразуется в различные форматы визуализации, однако неудобна при наборе и корректировке, так как каждое слово приходится набирать или исправлять трижды. На этих стадиях мы используем так называемый «компактный синтаксис», содержащий наряду с тэгами XML специальные символы, которые позволяют автоматически преобразовывать текст в «расщепленный» формат. Аналогичный подход применяется в некоторых системах совместного редактирования веб-сайтов, например MediaWiki.

На Рис. 6 представлен фрагмент «компактной» транскрипции «Страсбургских клятв».

```
<lb/> {{P:1:black}}ro #((do&bar;_d[e]o)) amur &amp;
((&pflour;_p[ro]))+? ((xp&bar;_[christ]))i_an poblo &amp;
((nro&bar;_n[ost]ro)) c((o&bar;))mun &slong;al*uament
```

Рис. 6. Фрагмент «компактной» транскрипции «Страсбургских клятв»

Открывающая текст буквица помещена в двойные фигурные скобки, в которых двоеточиями разделены данные о ее размере и цвете. Следующее слово, *deo* ‘Бог’, в рукописи сокращено. Границы сокращения отмечены двойными скобками, сокращенная и полная формы разделены подчеркиванием, «восстановленная» буква помещена в квадратные скобки. Перед словом поставлен знак «решетка», указывающий на то, что строчная буква рукописи должна быть преобразована в прописную в нормализованной транскрипции.

В начале второй строки примера записанные в сокращенной форме слова *pro* и *christian* отделены друг от друга малым пробелом. В транскрипции на это указывает сочетание знаков «плюс» и «вопросительный знак». В последнем слове примера «звездочка» перед *u* показывает, что в нормализованной транскрипции эта буква должна быть заменена на *v*.

Состав корпуса

Подготовленный нами для исследования средневековой французской пунктуации корпус многоуровневых транскрипций включает 28 отрывков прозаических текстов объемом от 550 до 2250 текстоформ. Общий объем корпуса составляет около 28 100 текстоформ. XII век представлен шестью рукописями, а XIV век – пятью. XIII 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 системой типологической классификации. Важнейшим параметром типологического описания текста является его принадлежность к определенной функционально сфере. В нашем корпусе по 9 текстов относятся к литературной и к научно-дидактической сферам, 6 текстов – к исторической и 2 – 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.

Аналитическая разметка

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 мы различаем 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 «экстра-сильной» 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.

В некоторых рукописях наблюдается употребление букв, занимающих промежуточное положение между строчными и прописными. Это либо укрупненные строчные буквы, либо малые прописные. В этих случаях можно говорить о средней силе пунктуации. На материале нашего корпуса подобное явление встречается крайне редко, причем речь идет скорее о потере графического различия между строчными и прописными буквами, чем о сознательно вводимом «третьем элементе» оппозиции.

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).

В процессе аннотации пунктуационные границы размечались по следующим основным типам стыкующихся единиц:

1) единицы макроструктуры текста (главы, эпизоды и т.п.);
2) прямая речь (начало, конец, реплика, обращение и т.п.);
3) вводные конструкции;
4) автономные предикативные единицы;
5) предикативные единицы с общими компонентами;
6) придаточные предложения;
7) однородные синтагмы, перечисление;
8) обособленные синтагмы (приложения, обстоятельства и т.п.);
9) прочие случаи (употребление знаков препинания вне основных пунктуационных границ).

Следует обратить внимание на отсутствие в классификации такой синтаксической единицы, как сложное предложение. Это связано с тем, что достаточно сложно на основании объективных, не зависимых от пунктуации критериев отличить независимые предложения от частей сложного с сочинительной или бессоюзной связью. С большей степенью объективности можно определить состав предикативных единиц (простых предложений), а затем проанализировать связь между этими единицами. Наиболее тесная связь образуется между главным и придаточным предложением (которое занимает позицию члена главного предложения). Не связанные отношениями подчинения предложения могут иметь общие члены (подлежащее или какой-либо из второстепенных членов). Для двух однородных придаточных предложений «общим членом» является главное. При отсутствии подчинения или общих членов предикативные единицы считаются автономными, даже если тесная семантическая связь между ними представляется очевидной.

Эксплуатация корпуса

По завершении разметки корпус загружается в поисково-аналитическую машину Weblex 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 (Textométrie), 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. В настоящее время подготовлен прототип платформы TXM, который наряду с Weblex будет продемонстрирован в действии на практическом занятии.

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

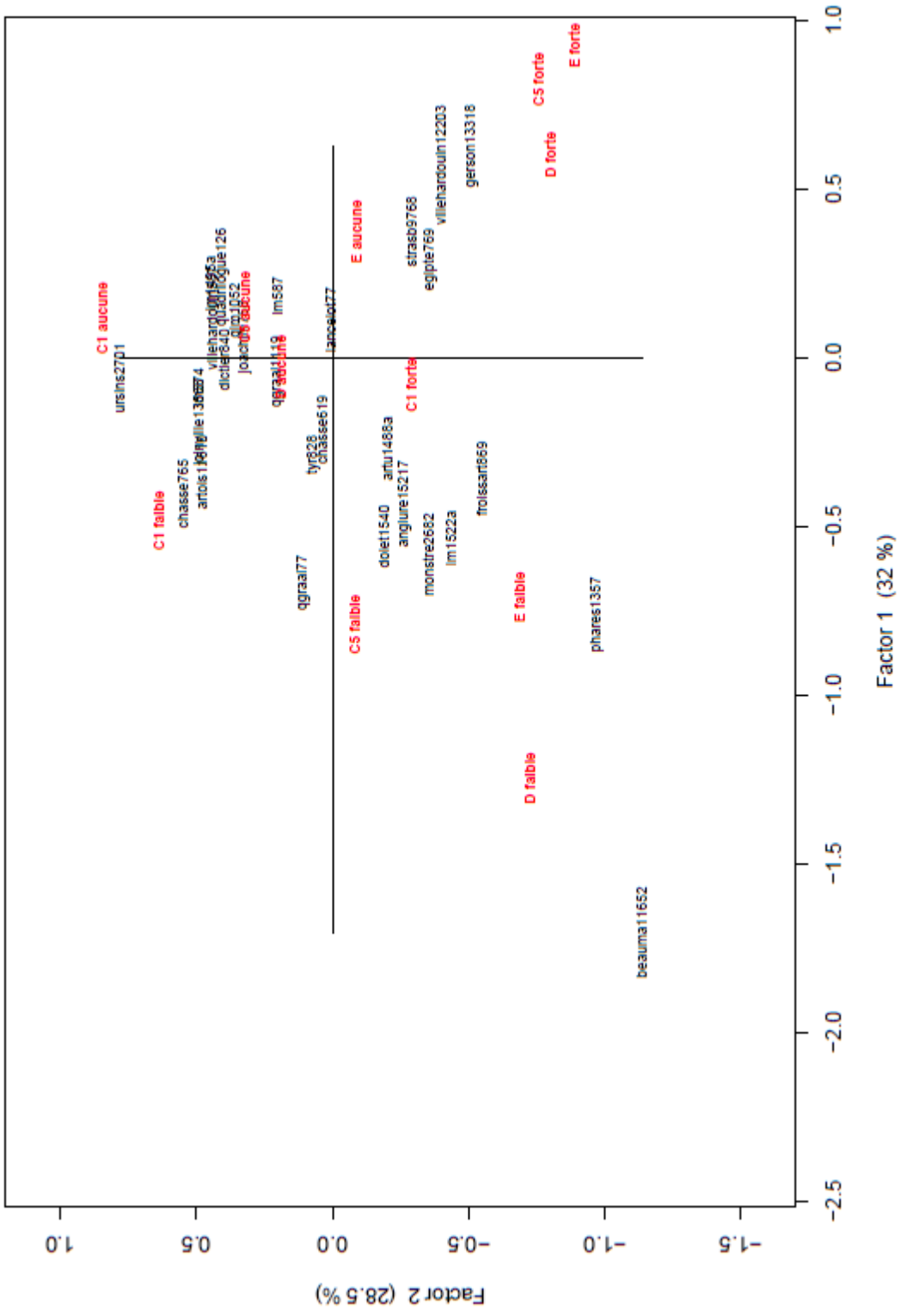

Рис. 7. Факторный анализ «силы» пунктуации на различных типах границ

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.

Полученный факторный план представлен на Рис. 7.

На данном графике мы можем наблюдать, с одной стороны, сходства в пунктуационном оформлении определенных типов границ, а с другой – группировки текстов в соответствии с отмеченными в них тенденциями пунктуации.

На графике хорошо видно, 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 Э. Доле «О правильной манере перевода» (1540) и «Страсбургские клятвы». 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.

Результаты исследования пунктуации

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%. Встречаются и исключения: в двух рукописях общий уровень пунктуации достигает 13%, а в одной рукописи, напротив, не превышает 3%. Заметим, что в современных французских текстах этот уровень достигает 14 – 16%.

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

Другим анализируемым параметром является относительная частотность сильной и слабой пунктуации. В 10 рукописях сильная и слабая пунктуация употребляются сбалансированно. В одной рукописи (XV в., повествование о крестовом походе) употребляется исключительно сильная пунктуация (в основном – прописная буква без знака препинания). В 7 рукописях сильная пунктуация доминирует (70 – 80% употреблений). В 5 рукописях и в печатных книгах в тех же пропорциях доминирует слабая пунктуация, и, наконец, в одной рукописи (XIV в., хроника Жана Фруассара) слабая пунктуация используется в 95% случаев.

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 4 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. «Комма» используется почти исключительно в слабой пунктуации и, возможно, соответствует определенной интонационной конструкции (с повышением тона) [Marchello-Nizia 2007]. Вместе с тем «доминирующий» знак может использоваться вместо «специализированного» во всех позициях.

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.

Литература

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.

Харман, Г. Современный факторный анализ: [Пер. с англ.] – М.: Статистика, 1972.

Andrieux-Reix, N., Monsonégo, S. Écrire les phrases au Moyen Âge. Matériaux et premières réflexions pour une étude des segments graphiques observés dans des manuscrits français médiévaux // Romania. – 1997. – vol. 115, № 459-460. – C. 289-336.

Baddeley, S. La ponctuation de manuscrits français du IX^{e} au XII^{e} siècle // Liaisons HÉSO/AIROÉ. – 2001. – № 32-33. – C. 139-149.

Barbance, C. La ponctuation médiévale: quelques remarques sur cinq manuscrits du début du XV^e^ siècle // Romania. – 1992-1995 . – vol. 113, № 455-456. – C. 505-525.

Buridant, C. Le strument *et* et ses rapports avec la ponctuation dans quelques textes médiévaux // Théories linguistiques et traditions grammaticales / éd. Anne-Marie Dessaux-Berthonneau. – Villeneuve-d'Asq: Presses Universitaires de Lille. – 1980. – C. 13-53.

Haugen, O. E. Parallel Views: Multi-level Encoding of Medieval Nordic Primary Sources // Literary and Linguistic Computing. – 2004. – vol. 19, № 1. – C. 73-91.

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.

Lavrentiev, A. Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIII^e^ au XV^e^ siècle: thèse de Doctorat non publiée. – Lyon: ENS LSH, 2009.

Li, H.-C. Découpage et structuration du texte: Lettrines, Majuscules, Blancs et autres signes de ponctuation dans les versions manuscrites et imprimée du Roman de Perceforest: Étude comparative: thèse de Doctorat non publiée. – Strasbourg: Université Marc Bloch – Strasbourg 2, 2007.

Llamas Pombo, E. Escritura y oralidad en los *Ovidiana* franceses des siglo XII: thèse de Doctorat non publiée. – Salamanca: Universidad de Salamanca, 1996.

Marchello-Nizia, C. Ponctuation et "unités de lecture" dans les manuscrits médiévaux ou: je ponctue, tu lis, il théorise // Langue française. – 1978. – № 40. – C. 32-44.

Marchello-Nizia, C. Le *comma* dans un manuscrit en prose du 13^e^ siècle: grammaticalisation d'un marqueur de corrélation, ou marquage d'intonation ? // Discours, diachronie, stylistique du français. Études en hommage à Bernard Combettes / éd. Olivier Bertrand *et al*. – Berne: Peter Lang, 2007. – C. 293-305.

Mazziotta, N. Ponctuation et syntaxe dans la langue française médiévale: Étude d'un corpus de chartes originales écrites à Liège entre 1236 et 1291: thèse de Doctorat non publiée. – Liège: Université de Liège, 2007.

Parkes, M. B. Pause and effect: an introduction to the history of punctuation in the West. – Aldershot : Scolar Press, 1992.

Roques, M. Le manuscrit fr. 794 de la Bibliothèque Nationale et le scribe Guiot // Romania. 1952. – vol. 73. – C. 177-199.

Textes d'étude (Ancien et moyen français) / éd. R.-L. Wagner, O. Collet. – Genève : Droz, 1995.