

ПРИНЦИПЫ ФОРМИРОВАНИЯ СЛОВАРЕЙ ДЛЯ ДЕШИФРОВКИ ЗНАМЕННЫХ ПЕСНОПЕНИЙ¹

Бахмутова Ирина Владимировна, Гусев Владимир Дмитриевич,
Титкова Татьяна Николаевна²

Аннотация

Рассматривается нерешенная в общем случае задача перевода древнерусских знаменных песнопений в современную нотолинейную форму. Ввиду контекстной зависимости языка знаменного распева соответствие "знамя – нота" является многозначным. Поэтому в качестве единиц дешифровки желательно выбирать не отдельные знамена, а более крупные структурные единицы с меньшей степенью неоднозначности. Предлагаются различные критерии и алгоритмы выделения таких единиц из текстов песнопений, представленных параллельно в знаменной и нотолинейной форме.

Введение

Древнерусские церковные песнопения XII – XVII веков в большинстве своем представлены в знаменной форме записи. Знамена – специальные знаки, служащие для передачи музыкальных звуков. Они интерпретируются как цепочки нот переменной длины (обычно от 1 до 5 нотных знаков). Процесс перевода песнопений из знаменной формы записи в нотолинейную, называемый *дешифровкой*, в общем случае не формализован, в силу чего певческие рукописи XVI века и более раннего периода практически нечитаемы. Известны лишь отдельные примеры дешифровки *пометных* рукописей XVII века (см., например, [1]), где процесс дешифровки был сильно облегчен наличием у знамен степенных и указательных помет, определяющих звуковысотную привязку и характер исполнения отдельного знамени.

В основе указанных дешифровок лежат азбуки "знамя – нота", созданные в конце XIX века известными знатоками знаменного распева – В.М. Металловым, Д.В. Разумовским и С.В. Смоленским, а также сборники *попевок* – элементарных интонационных единиц знаменного распева. Азбуки и особенно кокизники (сборники попевок) обладают разной степенью полноты, порою противоречивы, представляют не весь спектр структурных единиц и малопригодны для дешифровки беспометных рукописей ввиду отсутствия звуковысотных привязок. В связи с этим актуальным представляется формирование новых (машинных)



¹ Работа выполнена в рамках проекта №96–04–06258, поддержанного РГНФ, опубликована в: Труды 4-й Всероссийской с международным участием конференции "Распознавание образов и анализ изображений: Новые информационные технологии", Новосибирск, 11–18 октября 1998 г., с. 37–41.

² Лаборатория анализа данных, Институт математики им. С.Л.Соболева Сибирского отделения РАН. 630090 г.Новосибирск пр.Коптюга 4. E-mail: gusev@math.nsc.ru

словарей непосредственно по текстам песнопений, что и является целью данной работы.

1. Обоснование подхода

Основная трудность дешифровки, на наш взгляд, связана с тем, что язык знаменного распева *контекстно зависим*. Это проявляется в том, что интерпретация отдельных знамен зависит от типа элементарных структурных единиц, в которые они входят, расположения этих единиц в мелодии, гласовой принадлежности³ и ряда других факторов, что обуславливает многозначность соответствия "знамя – нота" и "нота – знамя".

Например, знамя  ("статья с запятой"), дважды входящее в состав попевки "кавычка" из гласа б: , в третьей позиции интерпретируется цепочкой из трех нот *e2d2c4*, а в пятой позиции – одной (целой) нотой *d1* (здесь и далее звуки первой октавы кодируются заглавными буквами: *G* – "соль", *A* – "ля", *H* – "си", звуки второй – прописными: *c* – "до", *d* – "ре", *e* – "ми", *f* – "фа", *g* – "соль", *a* – "ля", *b* – "си-бемоль", а цифры справа указывают длительность ноты:

1 – "целая", 2 – "половинная", 4 – "четвертная" и т.д.).

Ввиду контекстной зависимости языка целесообразно использовать в качестве единиц дешифровки не отдельные знамена, характеризующиеся высокой степенью многозначности, а более крупные структурные единицы ("знамена в контексте") с минимально возможной степенью неоднозначности.

В основе предлагаемого подхода лежит:

1) использование *двознаменников* для формирования словарей структурных единиц. Двознаменники – это билингвы знаменного распева, т.е. тексты, представленные параллельно в знаменной и нотной форме записи;

2) автоматическое *выделение* структурных единиц из *слитного* предварительно нерасчлененного *знаменного* текста. Нотный текст и последовательность помет (если они проставлены) могут облегчить решение этой задачи, но не всегда нами используются, поскольку в общем случае они отсутствуют. Двознаменник, как таковой, необходим лишь на этапе интерпретации выделенной структурной единицы.

Задача *выделения структурных единиц из слитного текста* сводится к выработке набора критериев, формализующих понятие "элементарная семантическая единица", и разработке эффективных алгоритмов, реализующих эти критерии. Часть критериев носит общелингвистический характер (частотные – позиционные критерии), другая – учитывает специфику знаменного распева. Ниже они рассмотрены более подробно.

³ В древних песнопениях понятие гласа ассоциировалось с ладом. Певческие книги разбивались на 8 частей (гласов), отличительные особенности которых регламентировались системой "осмогласия". В более позднее время произошло "размывание" ладовых систем и понятие гласа стало ассоциироваться с совокупностью функционирующих в нём попевок.

2. Критерии выделения структурных единиц.

2.1. *Частотный критерий* [2] основан на предположении, что элементарные семантические единицы языка проявляют себя в виде устойчиво повторяющихся цепочек символов. Понятие устойчивости нуждается в уточнении.

Пусть $p = a_1 a_2 \dots a_n$ – произвольная цепочка (слово) текста T , составленная из элементов алфавита Σ . Для всевозможных префиксных подслов этого слова имеют место следующие соотношения:

$F(a_1) \geq F(a_1 a_2) \geq \dots \geq F(a_1 a_2 \dots a_{n-1}) \geq F(p)$, где $F(\alpha)$ – частота встречаемости слова α в тексте T . Назовем цепочку p устойчивой при правостороннем расширении (или устойчивой справа), если:

1) существует $1 \leq i < n$, такое что $F(a_1 a_2 \dots a_i) \succ F(a_1 a_2 \dots a_i a_{i+1}) \succ \dots \succ F(p)$, где знак " \succ " означает "не намного больше" (это условие появления доминантной по частоте линии);

2) $F(p) > 1$;

3) $F(p a_{n+1}) = F(a_1 a_2 \dots a_n a_{n+1}) \ll F(p)$, где $a_{n+1} \in \Sigma$, а $p a_{n+1}$ – произвольная цепочка текста T .

Первое условие означает, что как только формируемая цепочка приобретает содержательный смысл, ее продолжение легко предсказуемо, что приводит к стабилизации частот (слово "чемпион", например, предсказуемо уже по цепочке "чемп"). Условие 2 исключает все цепочки с $F=1$, поскольку любые их расширения имеют ту же частоту, и формально являются неделимыми. Условие 3 соответствует прерыванию доминантной по частоте линии, фиксируемой условиями 1 и 2 (слово "чемпион" продолжаемо уже разными способами: чемпионат, чемпионский, чемпионка и т.п.).


Аналогично цепочку $p = a_1 a_2 \dots a_n$ текста T назовём устойчивой при левостороннем расширении (устойчивой слева), если существует $1 \leq j \leq n$ такое, что:

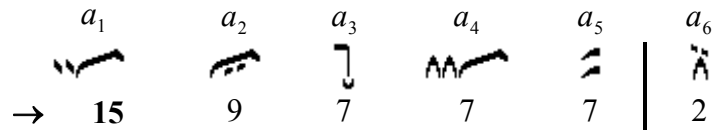
$$F(a_j a_{j+1} \dots a_n) \succ F(a_{j-1} a_j \dots a_n) \succ \dots \succ F(p), \quad F(p) > 1,$$

$F(a_0 p) = F(a_0 a_1 \dots a_n) \ll F(p)$, где $a_0 \in \Sigma$, а $a_0 p$ – произвольная цепочка текста T . Потенциальной *структурной единицей* будем считать цепочку текста, устойчивую как слева, так и справа. Это фрагмент, заключенный между левой и правой точками прерывания доминантной по частоте линии.

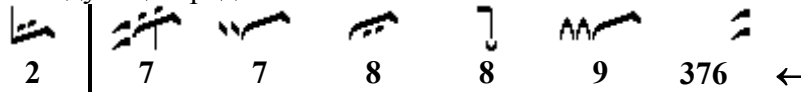
Алгоритм отыскания структурных единиц напоминает "качели". Отправляясь от какого-либо элемента алфавита (или цепочки элементов), анализируем возможные продолжения в поисках участков стабилизации частот. Зафиксировав правую границу, идем в обратном направлении в поисках левой границы. Для получения частот всевозможных подцепочек предварительно вычисляем полный частотный спектр l -грамм текста $\Phi(T) = \{\Phi_l(T)\}$, где $\Phi_l(T)$ – частотная характеристика l -го порядка, $1 \leq l \leq l_{\max}$, l_{\max} – длина максимального повтора в тексте. $\Phi_l(T)$ содержит информацию о всех повторах длины l ,









содержащихся в T . Здесь $T = T_1 * T_2 * \dots * T_m$ – конкатенация песнопений одного гласа, m – число песнопений, $*$ – разделитель, а рассматриваются лишь повторы, не содержащие $*$. Для вычисления $\Phi(T)$ используется процедура рекуррентного хеширования [3].

Пример. Знамя  ("тряска") встречается 15 раз в песнопениях гласа 4 из певческой книги "Октоих" XVII века (собрание Кирилло – Белозерского монастыря). Рассмотрим доминирующие по частоте правосторонние расширения этого знамени и проследим за изменениями частот $F(a_1)$, $F(a_1a_2)$ и т.д.:










Здесь число, стоящее в нижней строке под символом a_i , означает $F(a_1, a_2 \dots a_i)$, $i = 1, 2, \dots$. Нетрудно видеть, что участок стабилизации частот обрывается при переходе от a_5 к a_6 . Ставим между ними правую границу (|) и начинаем с a_5 движение влево, отслеживая частоты $F(a_5)$, $F(a_4a_5)$, $F(a_3a_4a_5)$ и т.д. Получаем следующий ряд чисел:



Для выявления границы участка стабилизации нам пришлось продвинуться на два шага левее исходного знамени. Левая граница проходит между знаменами  и . Потенциальная семантическая единица – фрагмент, заключенный между левой и правой границами:      . Это попевка – производная от архетипа "долинка".



Описанный подход хорошо выделяет наиболее массовые попевки гласа, а также некоторые структуры, не относящиеся к категории попевок. Однако он дает сбой на коротких малочастотных попевках, где отсутствует зона стабилизации частот и теряют смысл понятия "много больше" ($>>$), "не намного больше" ($>$) и т.п.

2.2. *Частотно – позиционный критерий* [4] устраняет некоторые недостатки частотного и существенно ускоряет поиск структурных единиц. Идея подхода сводится к тому, что, используя некоторую априорную информацию о знаменах, а также взаимосвязь стихотворного и знаменного текстов, удастся выявить множество знамен, *завершающих* попевки (, , , , ,  и др.). В результате процедура правостороннего расширения становится ненужной. В тексте выявляются поочередно все вхождения каждого из кадансовых (завершающих) знамен. Предшествующие им цепочки, прочитываемые справа – налево, упаковываются в лексикографическое дерево со склеенными общими началами (они соответствуют концам попевок). Корнем дерева является кадансовое знамя. Левые концы попевок фиксируются как границы участков стабилизации частот при движении от корня к листьям. *Низкочастотные* (и, как правило, короткие) цепочки












деревя могут быть идентифицированы по аналогии с соседними высокочастотными. Недостатками этого подхода являются: возможность пропуска кадансового знамени на этапе формирования множества знамен, завершающих попевки, возможность появления кадансового знамени не в завершающей части попевки (см. знамя  в третьей позиции попевки "кавычка", п.1) и ориентированность всего подхода лишь на структуры типа "попевка".






2.3. *Метод "всплывающего пузырька"*. Еще более упрощенным аналогом частотного критерия является метод, основанный на наблюдении за рангом последовательно удлиняющейся цепочки в частотных упорядочениях всевозможных подцепочек текста той же длины. Если ранг вначале уменьшается (т.е. цепочка поднимается в упорядочении), затем стабилизируется, а потом начинает нарастать (часто скачком), то весьма вероятно, что мы имеем дело со структурной единицей, длина которой определяется по состоянию, предшествующему скачку.

2.4. *Выявление звуковысотных инвариантов*. В отличие от предыдущих подходов этот принципиально требует наличия двознаменника. Анализ частотных характеристик $\Phi_l(T), 1 \leq l \leq l_{\max}$, показывает, что многозначность соответствия "знамя – нота" не носит абсолютного характера. Отдельные знамена (а чаще цепочки знамен) допускают в пределах одного гласа (а иногда и нескольких) однозначную интерпретацию. Они могут служить *звуковысотными ориентирами* при дешифровке *беспометных* рукописей.

Так, анализ двознаменного "Октоиха" XVII века (Соловецкое собрание) показал, что в гласах 1, 2, 4, 6 однозначно интерпретируется "стрела светлая" () — код $e4f4g2$, в гласах 3,4 – "дербица": () – код $H4c4d4e4$ и т.д. Свойство однозначной интерпретируемости свидетельствует о высокой *информативности* соответствующих знамен (или цепочек), поскольку во всех языковых системах фрагменты текста, не меняющиеся в ходе эволюции, считаются наиболее значимыми в функциональном отношении. Число однозначно интерпретируемых цепочек возрастает с увеличением их длины. Недостатком подхода является игнорирование малочастотных, хотя и однозначно интерпретируемых цепочек, поскольку их "однозначность" может быть случайным фактором (следствием низкочастотности). Специального обоснования требует и перенос "по аналогии" свойства "однозначной интерпретируемости" на беспометные рукописи. Это обоснование связано с анализом эволюции знаменного распева.

2.5. *Тандемные повторы*, часто встречающиеся в знаменных песнопениях, также можно отнести к классу структурных единиц, не отраженных в традиционно используемых дешифровочных таблицах. Они требуют специального изучения, поскольку далеко не всегда повторяемые фрагменты интерпретируются одинаково:

										
$a1$	$g2$	$g4f4$	$g4a4$	$b2$	$a2$	$g2$	$g2e4$	$f4g4$	$a2$	$g2$
	ру-	ко-	пи-	са-	ни-	е	при-	гво-	зди	на

2.6. *Гласоспецифичные цепочки* характеризуются тем, что встречаются лишь в отдельных гласах. Так, знамя  ("крюк светлый с сорочьей ногой") встречается только в гласе 3 двознаменника в составе устойчиво повторяющейся конфигурации     ($F = 6$), обычно открывающей внутренние разделы песнопений. Гласоспецифичные цепочки информативны в силу своей уникальности. Включение их в дешифровочные словари позволит точнее отразить специфику гласов и выявить нередкие случаи перетранспонирования гласов по высоте.

Заключение

Предложены различные критерии и алгоритмы для выделения структурных единиц из текстов знаменных песнопений. Они могут частично пересекаться, но в общем случае взаимно дополняют друг друга. Наибольший вес имеют структурные единицы, выделяемые по совокупности критериев. Предполагается, что их использование в качестве единиц дешифровки приведет к существенному снижению неоднозначности по сравнению со случаем дешифровки по знаменам.

СПИСОК ЛИТЕРАТУРЫ

1. Бахмутова И.В., Гусев В.Д., Титкова Т.Н. Компьютерный анализ древнерусских двознаменников: многозначность соответствий "знамя – нота" и "нота – знамя". // Искусственный интеллект и экспертные системы. – Новосибирск, 1996. Вып. 157: Вычислительные системы. – С. 68–100.
2. Бахмутова И.В., Гусев В.Д., Титкова Т.Н., Шиндин Б.А. Дешифровочный подход к анализу древнерусских песнопений // Анализ последовательностей и таблиц данных. – Новосибирск, 1994. – Вып. 150: Вычислительные системы. – С. 107–130.
3. Гусев В.Д., Титкова Т.Н. Рекуррентное хеширование символьных цепочек. // Там же, С. 94–106.
4. Бахмутова И.В., Гусев В.Д., Титкова Т.Н., Шиндин Б.А. Об одном подходе к проблеме дешифровки древнерусских песнопений в невменной записи // Труды Сибирской конференции по прикладной и индустриальной математике, посвященной памяти Л.В. Канторовича. – Новосибирск, 1997. – Т. 2. – С. 1–10.