Применение методов машинного перевода для анализа древнерусских музыкальных рукописей

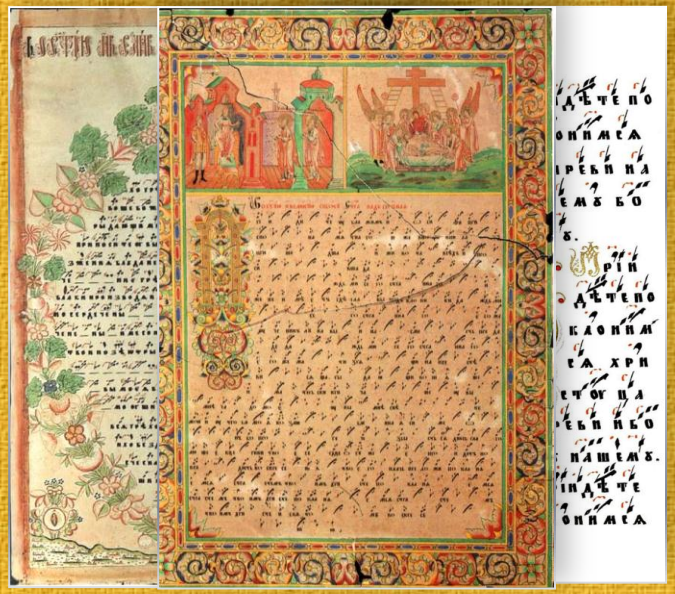
Даньшина М.В.
Научный руководитель – Филиппович А.Ю.

Древние музыкальные рукописи

Собрание Д.В.Разумовского и В.Ф. Одоевского

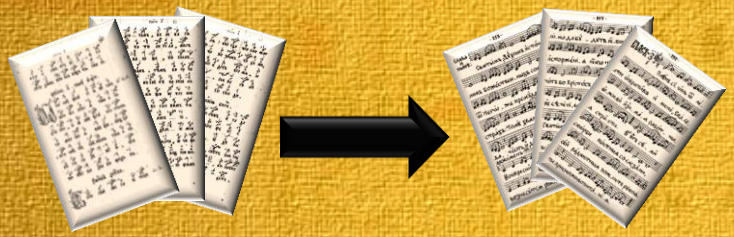
Сборники песнопений:

- на крюковых нотах – **83** рукописи
- на линейных нотах – **43** рукописи
- двузнаменные – **3** рукописи

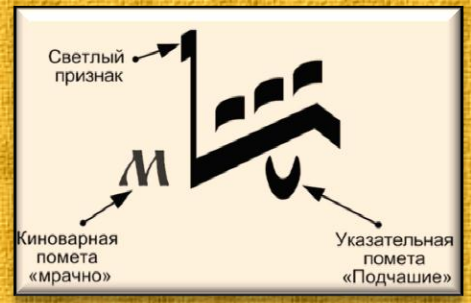


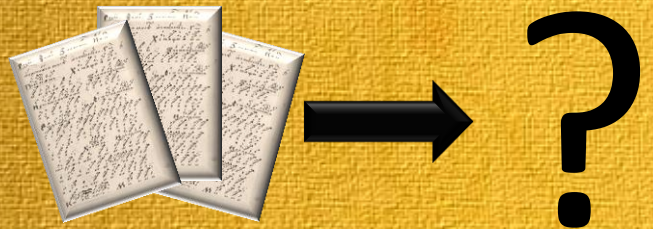
- Синодальное собрание славянских рукописей - **1172** фолианта.
- Рукописей Соловецкого монастыря - **202** рукописи
- Общество любителей древней письменности - **624** славяно-русских и нескольких греческих рукописей

Знаменные песнопения

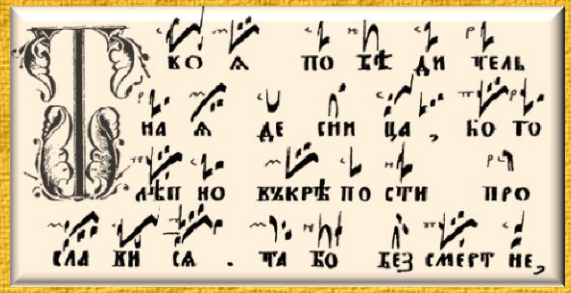


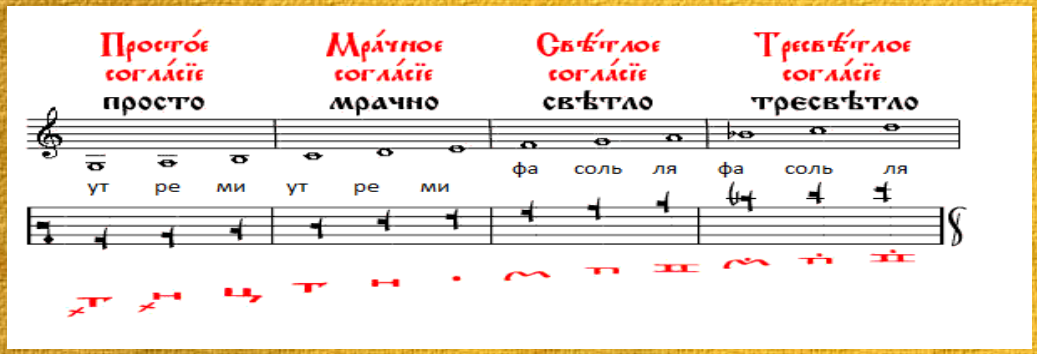
XVII-XVIII века



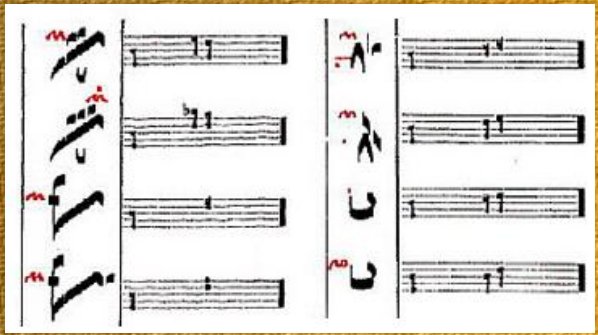


XI-XVI века





Сложность перевода



Знамя – графическое изображение ноты

Особенности знамен

- Перевод, представленный в азбуках, *неоднозначен*
- Одно знамя может представляться несколькими нотами
- Не все знамена представлены в азбуках

Попевка – последовательность знамен с особенные значением (переводом).

Особенности попевок

- Большое количество попевок **(>2 000)**
- Попевки объединялись в сборники
- Зависимость перевода от текста

Сборник попевок Соловецкого собрания содержит 213 страниц (~2000 попевок)

П={{Z},N,G}, где Z={<z,s>},
где П-попевка, N-название, G-глас, z-знамя, s-слог.

S={<z,nota,s>},
где S – словарь, nota-перевод знамени.

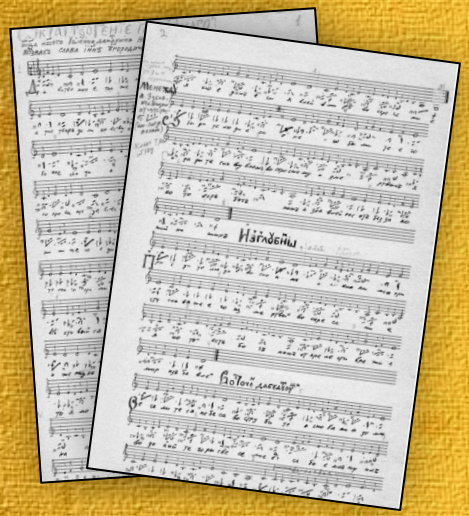
4

Исследуемые рукописи

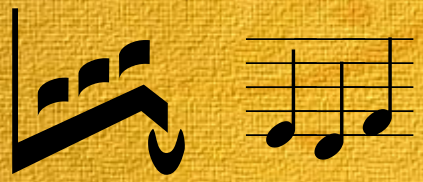
Круг древнего знаменного пения 1884г.

1367 страниц (6 томов)

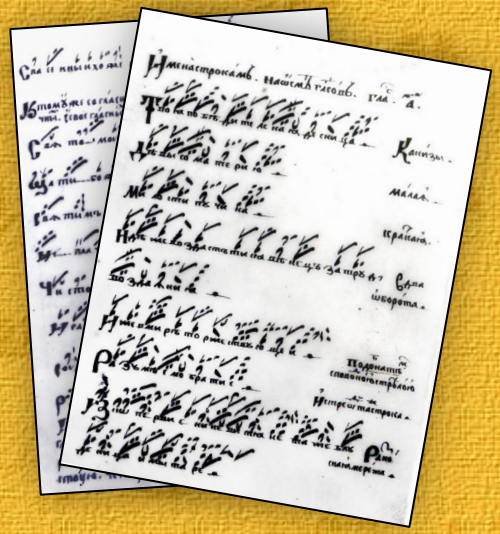
Ирмологий 17 в.



68 страниц



Сборник попевок Соловецкого собрания



213 страниц

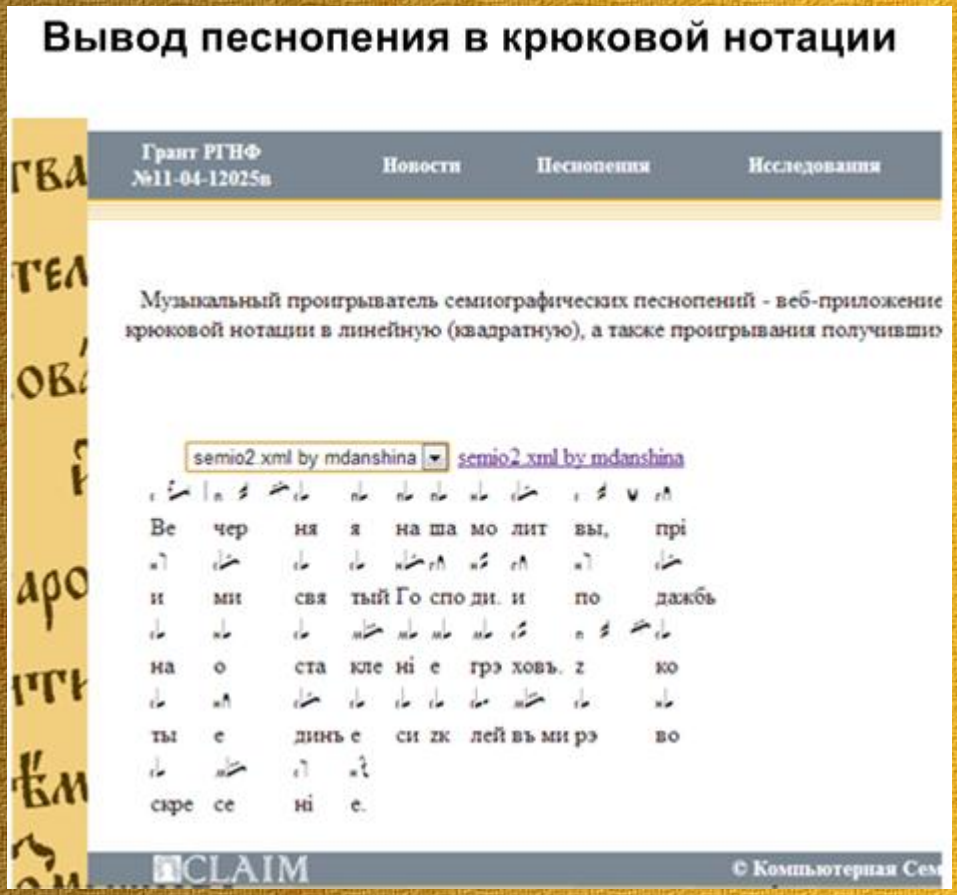
5

Использование методов машинного перевода

1) Прямой (пословный) перевод
2) Использование правил
3) Статистический перевод

Музыкальный редактор

Вывод песнопения в крюковой нотации



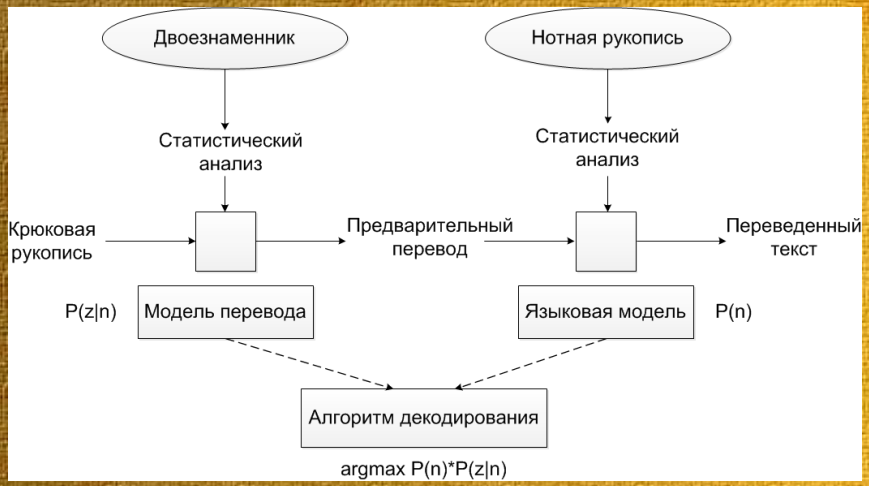
Пример азбуки в xml-формате

```
▼<ROWDATA>
   <ROW cod="200" note="e_1" length="04" p="0.7"/>
   <ROW cod="144" note="f_1" length="04" p="0.9"/>
   <ROW cod="304" note="g_1" length="04" p="0.9"/>
   <ROW cod="269" note="f_1" length="04" p="0.4"/>
   <ROW cod="177" note="e_1" length="04" p="0.4"/>
   <ROW cod="256" note="f 1" length="04" p="0.5"/>
```

Пример песнопений в xml-формате

```
▼<ROWDATA>
   <ROW Znam="a" Slog="ко" Stil=" обычный" VPom="м" DPom=""/>
   <ROW Znam="Ар" Slog="кэч" Stil=" обычный Italic" VPom="в" DPom=""/>
   <ROW Znam="a" Slog="но" Stil=" Bold" VPom="п" DPom=""/>
   <ROW Znam="a" Slog="му" Stil=" Bold" VPom="п" DPom=""/>
   <ROW Znam="a" Slog="t" Stil=" Bold" VPom="п" DPom=""/>
   <ROW Znam="a" Slog="от" Stil=" Bold" VPom="п" DPom=""/>
   <ROW Znam="a" Slog="шу" Stil=" Bold" VPom="п" DPom=""/>
   <ROW Znam="a" Slog="ся" Stil=" Italic" VPom="п" DPom=""/>
```

Статистический машинный перевод



Модель языка

$$P(w_3|w_1,w_2) =\cdot \frac{C(w_1,w_2,w_3)+1}{C(w_1,w_2)+V}$$

Модель перевода

$$P(n|z) = \frac{C(n,z)}{C(z)},$$

где C(n,z) – количество раз, когда последовательность знамен z переводится нотами n

Алгоритм декодирования

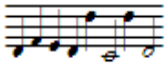
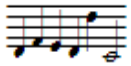
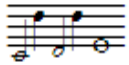
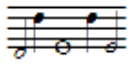
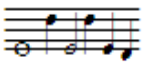
$$\arg\max_z p(z|n) = \arg\max_z p(z)\cdot p(n|z),$$

где z – триграмма знамен, n – перевод триграммы (ноты), p(z|n) — условная вероятность того, что переводу n соответствовал исходный фрагмент z

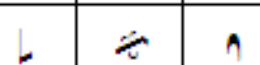
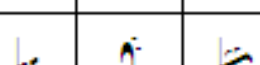
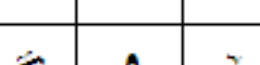
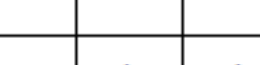
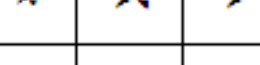
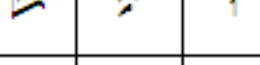
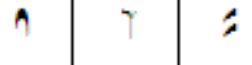
$$P(w_1,w_2,\dots,w_m) \cong \prod_{i=1}^{m} P(w_i|w_{i-n+1},w_{i-n+2}\dots w_{i-1}), n>1$$

$$P(w_i|w_{i-1},w_{i-2}) = \frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})}$$

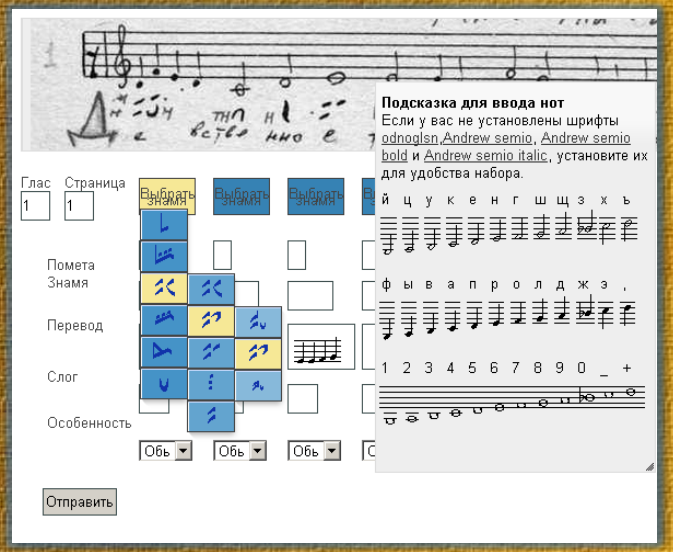
Пример модели языка

N-грамма	Вероятность со сглаживанием	Вероятность без сглаживания	N
	0,01139	0,666667	3
	0,015909	0,857143	2
	0,00431	0,032258	3
	0,056338	0,22963	2
	0,004348	0,037037	3
	0,025271	0,04	2
	0,004228	0,025	3
	0,049878	0,102828	2

Пример модели перевода

Триграмма			Перевод			Вероятность
						0,017327
						0,318182
						0,014851
						0,272727
						0,073529
						0,3125
						0,153846
						0,111111
						0,333333
						0,2

Ввод песнопений в БД в программе IPSM

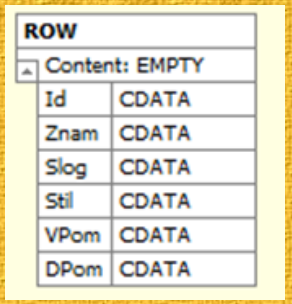


С помощью веб-приложения

- Для ввода нот предусмотрены специальные подсказки
- Текущая страница для ввода отображается на экране
- Сохранение дополнительной информации о песнопении (название, страница)

Загрузка XML-файла

- Добавление песнопений, набранных в MS Word
- Специальный формат XML-файла



Кодирование знамен

	2112300
	2113100
	2113200
	2113300

- Выделено 202 знамени
- Знамена разделены на 6 групп (максимум 7 подгрупп)
- От 10 до 66 знамен в группе

В результате БД содержит:

- 29376 записей из «Круга…»
- 234 записи приложения к «Круга..»
- 10897 записей двоезнаменника
- 16914 записей сборника попевок

Спасибо за внимание!