ГРАММАТИЧЕСКИЙ СЛОВАРЬ

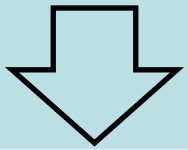
ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ XVIII—XIX в.: ПЕРВЫЕ РЕЗУЛЬТАТЫ

А.Е. Поляков, НПБ им. К.Д. Ушинского РАО
С.О. Савчук, Д.В. Сичинава, Институт русского языка им. В.В. Виноградова РАН

Мотивация

Отсутствие морфологического анализатора для исторических корпусов и электронных библиотек

Отсутствие анализатора для распознавания отсканированного печатного текста в орфографии XVIII—XIX вв.



Неудовлетворительные результаты поиска и низкое качество распознавания

Цели

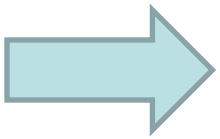
- Добиться удовлетворительных результатов при автоматическом анализе текстов следующих категорий.
- I. Тексты к. XVIII - н. XX в. в дореволюционной орфографии из оригинальных источников, не издававшиеся в XX в.
- *Середа. Маїя 5 дня 1787 года. Устъ не имѣя говорю: гласъ мой слушаютъ въ отдаленности и многїе оному повинуются. Иногда размѣряю и теченїе времени: иногда созываю людей на работу.*
- Основная проблема – правила графики и орфографии

- II. Тексты XVIII в. в современной и оригинальной орфографии.
- *Младыи* *человѣкЪ* *всегда* *имѣетЪ* *сЪ* *благочестными* *и* *добродѣтелными* *людми* *обходїтїся*, *отЪ* *которыхЪ* *бы* *онЪ* *добру* *научїтца* *могЪ*. *Также* *и* *сЪ* *такїми* *людми* *которые* *и* *честное* *имя* *и* *непорочное* *жїтїе* *имѣютЪ*. *А* *отЪ* *такїхЪ*, *которые* *легкомысленно* *и* *злочестно* *живутЪ*, *бѣгать* *яко* *бы* *отЪ* *яду* *или* *лютого* *мору*. [**Юности честное зерцало, 1717**]
- Основная проблема – пополнение словаря + исторические парадигмы + правила графики и орфографии

- III. Тексты XIV-XVII в. и рукописные тексты XVIII в.
- *И не точію тѣхъ иже градъ враждебно, яко зміеве, своими зубы держа щихъ, носихъ множество и своевѣрныхъ враждующихъ ми страшахся, иже присѣдятъ о насъ тайно въ ловителѣхъ ко Еллиномъ* [**Иван Тимофеев. Временник (1610-1617)**]
- *и пажалои ты приежьщаи х празнику Никалаю чюдворцу а ка мнѣ въ госте а я млсте твоеи вседошна ратъ писавы Івашка Сидарав челомъ бьетъ* [**И. Сидоров И.А. Снарскому (1701)**]
- Основная проблема – пополнение словаря + исторические парадигмы + отсутствие норм орфографии

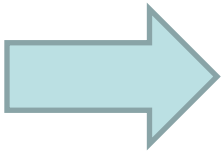
Задачи

- 1. *Грамматические*: разработка грамматических парадигм для лексем, отсутствующих в современном словаре.



- Грамматический словарь

- 2. *Орфографические*: обеспечение лемматизации форм, имеющих отклонения от стандартных написаний.



- Правила

Грамматический словарь

список лексем с приписанной информацией о словоизменении.

1) основа с указанием чередований	2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность)	3) код словоизменительного типа (парадигмы)
б(и|ь|е)+ть V,ipf,tr V11	пе(к|ч|)+ь V,ipf,tr V8	но(с|ш)+ить V,ipf,tr V4

Модули грамматического словаря

Современный модуль

- Грамматический словарь Зализняка

Модуль XVIII—XIX века

- Исторические словари
- Корпус текстов

Модуль XVII века

- Корпус текстов
- Словарь русского языка XIV—XVII в

Формирование словника

Первое направление: расширение словника за счет включения в него материалов исторических словарей

- – Словарь Академии Российской (1789–1794);
- – Словарь церковнославянского и русского языка (ЦСРЯ) (1847);
- – Полный русский орфографический словарь (1898);
- – Словарь русского языка XVIII века.

Второе направление: пополнение словника словами, извлеченными из корпуса текстов XVIII-XIX в., которые не учтены в лингвистических источниках

- Несловарные слова:
- Имена собственные и образованные от них прилагательные.
- Лексемы, отсутствующие в словаре (*ажиатёр*, *анатомить*, *бомбаст*)
- Морфологические архаизмы (*гресех, острови, знаеши, бяше, бяху* и пр.).
- Варианты различного рода.

Правила

А) *Графические правила*

- основаны на приравнивании графем, встретившихся в тексте, графемам, входящим в порождаемые словарём формы, например, *ѳ* => *ф; ѯ* => *кс*.

Б) *Орфографические правила*

- связаны с нормализацией определённых орфограмм. Например, *цы* => *ци*: *цыновка, цыдулка, цыгарка, цыгейка*

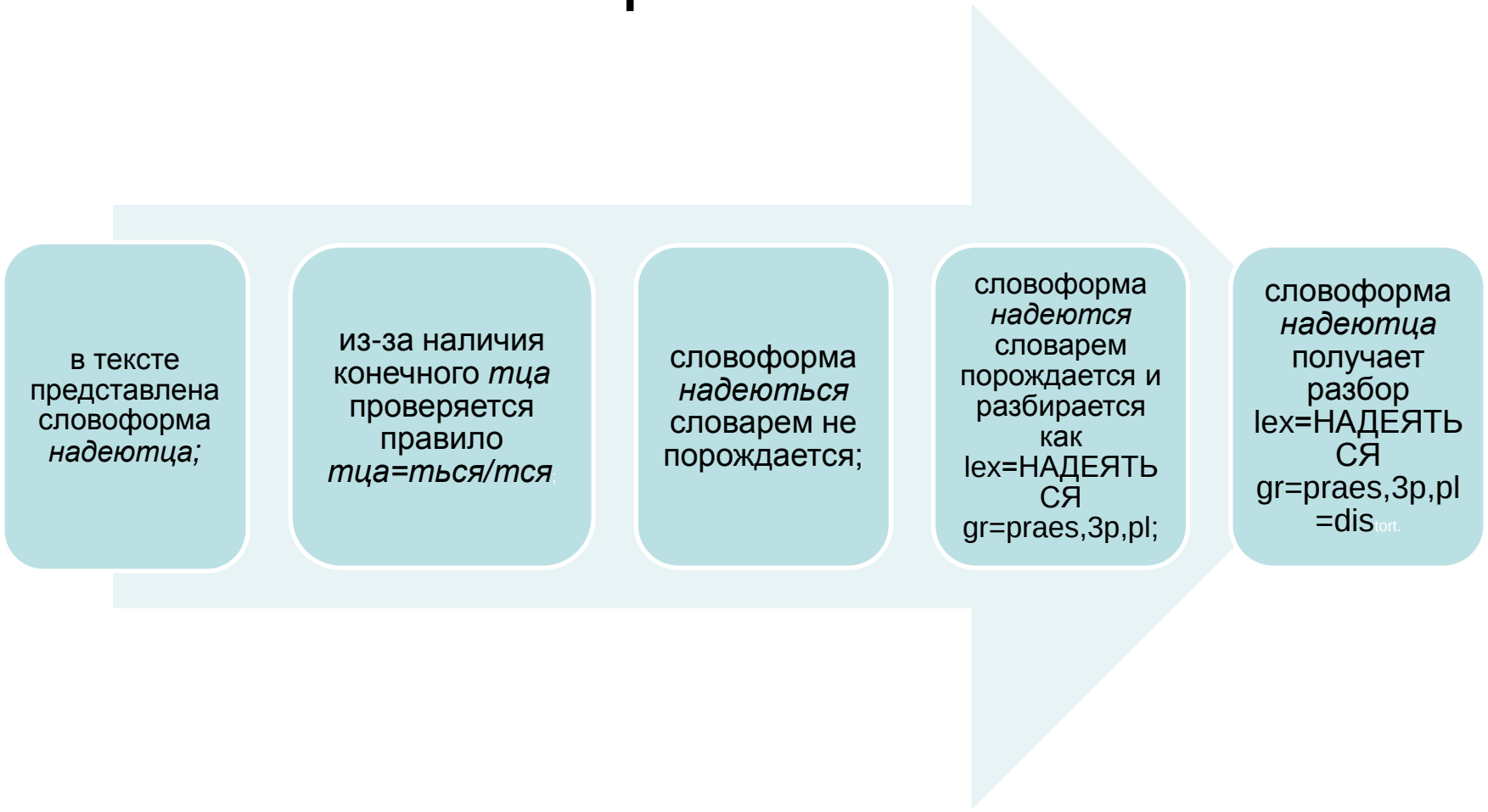
В) *Морфологические правила*

- устроены с учётом информации о грамматическом разборе словарной словоформы. Например, *-яе*, comp,anom <= *-ее,* comp.

Г) Списочный способ

- задание вариативности списками, на ограниченных классах единиц. Например, ряд конкретных корней в среднерусских и церковнославянских текстах записываться под титлом, ср. такие словоформы, как *Г(о)с(по)дь, м(е)с(я)ц, гл(агол)ет* и т. д.

Пример работы орфографического правила



Порядок применения правил

графические правила, в том числе списочные

↓

орфографические

↓

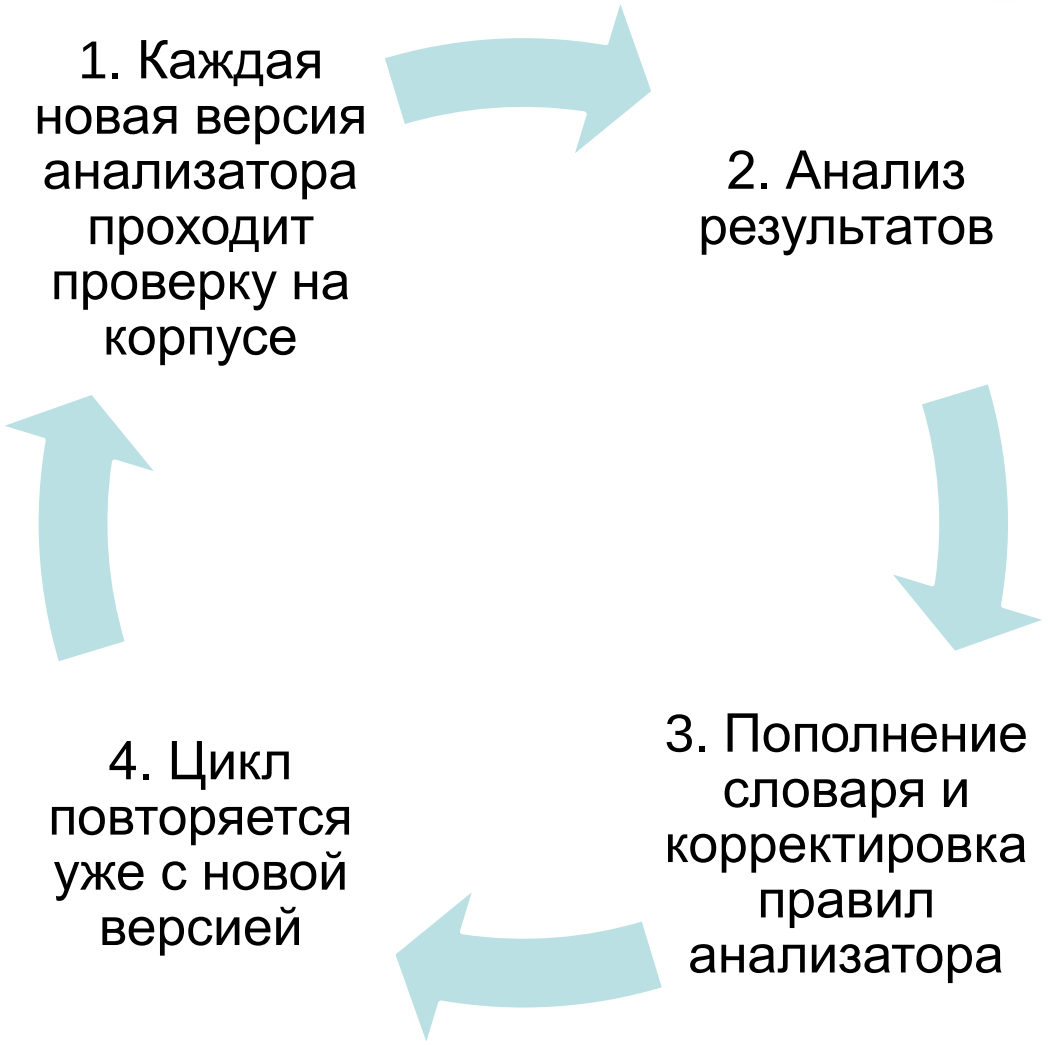
морфологические

↓

списочные правила, блокирующие паразитические разборы

Циклы обработки текстов

В процессе формирования словаря и оптимизации работы анализатора предусматривается несколько циклов обработки текстов.



Оценка результатов

Экспериментальный корпус –
4 млн с/у, ок. 256 тыс. разных с/форм

Состав текстов	
художественные	24%
публицистика	24%
церковно-богословские	19%
научные	17%
официальные	11%
бытовые	5%
Хронологическое распределение	
1700-1730	6%
1731-1780	43%
1781-1799	30%
1800-1830	21%

Не распознаны или неправильно опознаны

сочетания знаменательных частей речи **с частицами** *-то(-та, -ат), же(ж), ли(ль), бы(б), -де, -ка: таковаже, пили-б, пилиб, еслиб.*

архаические формы склонения и спряжения: *возвышаяй*, *живяше*, *знаяше, стояху*

орфографические варианты: *безщчетну, возмеш, баталиах*

собственные имена — топонимы, имена лиц, литературных героев и мифологических персонажей: *Шлюссембурх, Елисавефь, Невтон, Мишель Анжело*

разобрано	**185221**	**72,4%**
гипотезы	63904	25,0%
не разобрано	6780	2,6%

Оценка правильных разборов

Высокий показатель однозначных разборов (73%)

В корпусе мало текстов начала XVIII в.

Основной часть текстов представлена в современной орфографии

Увеличение доли оригинальных текстов XVII и 1-ой трети XVIIIв. приведет к росту количества ошибочных разборов

Ближайшие планы

- внедрение всех изменений в словарь;
- коррекция экспериментального корпуса с учетом выявленных ошибок;
- тестирование новой версии словаря на экспериментальном корпусе;
- тестирование словаря на отдельных текстах более раннего периода, относящихся к разным жанрам.

- Работа выполнена при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика».
- А.Е. Поляков pollex@mail.ru, НПБ им. К.Д. Ушинского РАО
- С.О. Савчук savsvetlana@mail.ru, Д.В. Сичинава mitrius@gmail.com Институт русского языка им. В.В. Виноградова РАН