Корпус древнерусских текстов «своими руками»: задачи и технологии

Пермь, 8 ноября 2019 г.

Дмитрий Анатольевич Добровольский,
доцент Школы исторических наук
ФГН НИУ ВШЭ (Москва)

ddobrowolski@hse.ru

Корпус — общегуманитарный инструмент

- Корпусный метод предполагает обработку больших объемов текста по определенным формальным признакам.
- Корпусный подход полезен при атрибуции исторических источников.
- Корпусный подход тесно связан с контент-анализом.

Следовательно,

- корпус нельзя рассматривать как сугубо лингвистический инструмент.

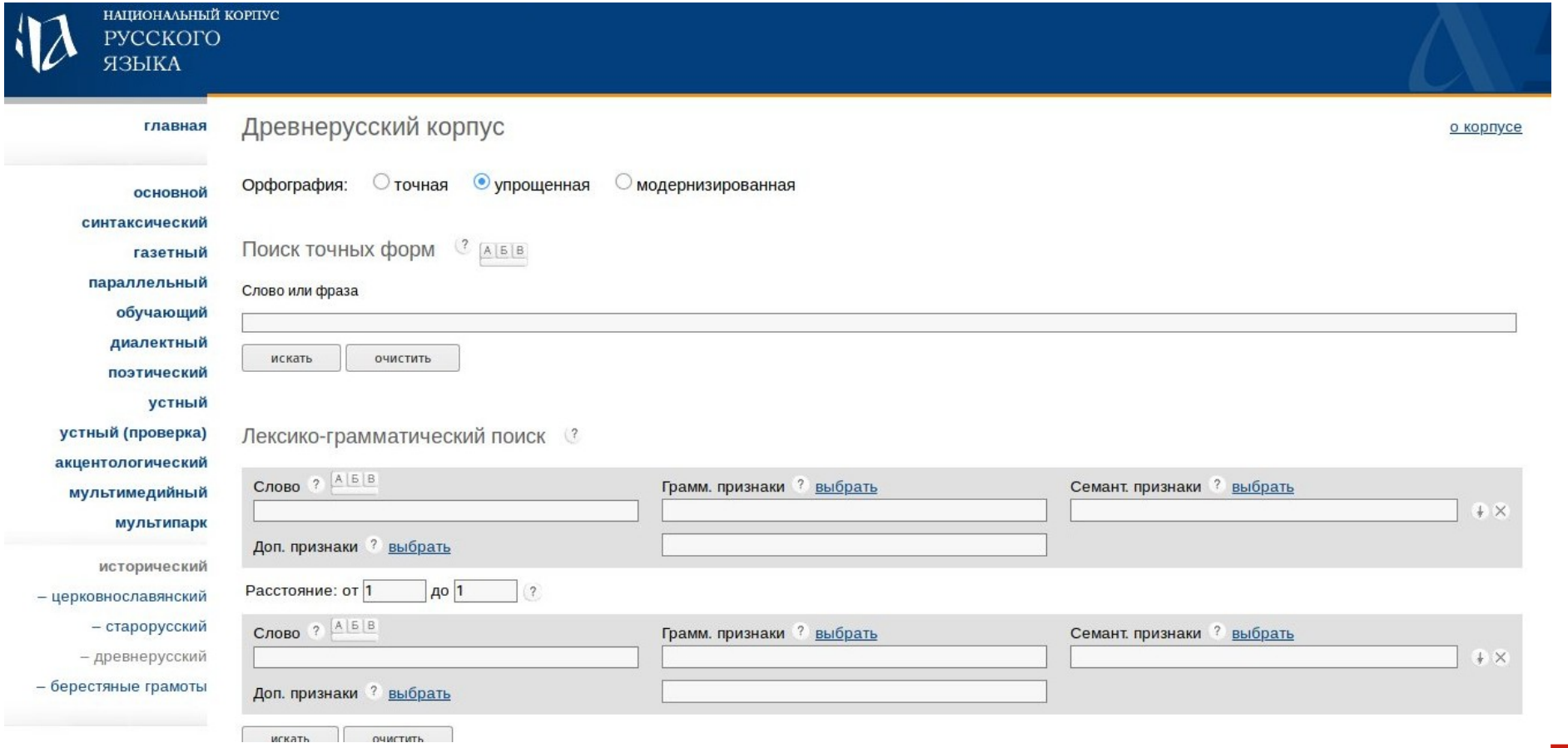
Пример задачи

- Феодосий Печерский († 1074) — не только один из основателей русского монашества, но и известный проповедник.
- Комплекс почти несомненных сочинений Феодосия сохранился в составе Торжественника РГБ, Рум. 406. (ср. известие Начальной летописи о кончине святого).
- Но традиция/историография приписывает Феодосию еще несколько произведений, в т.ч. полемическое «Слово о вере христианской и о латинской» и «Слово о казнях Божиих».

Лингвостилистический анализ

- В проповедях на третью неделю поста из Рум. 406, активно используется (а) частица **бо** как средство паратаксиса, (б) риторические вопросы с частицей **ли**.
- Для «Слова о вере христианской…» это не характерно; его синтаксис более замысловатый.
- В то же время: стечение частиц **бо** и **ли** обнаруживается в финальной (оригинальной) части «Слова о казнях Божиих».

Естественно...



Ограничения древнерусского подкорпуса НКРЯ

- Узкий круг источников.
- Данные закрыты, автоматические запросы затруднены.
- Закрытый лемматизатор и база прецедентов.

Плюсы:

- глубина проработки материала
- качество предлагаемых разборов

Минус:

- невозможно масштабирование под задачи большинства исследователей

```
@property
def lemma(self):
    return self._lemma

@lemma.setter
def lemma(self, value):
    self._lemma = value

@property
def part_of_speech(self):
    return self._part_of_speech

@part_of_speech.setter
def part_of_speech(self, value):
    valid_values = [
        None, 'N', 'ADJ', 'PRON', 'V', 'PTCP',
        'ADV', 'PRAE', 'CONJ', 'PTCL', 'INT'
    ]
    if value not in valid_values:
        raise ValueError(
            'unknown part of speech ("{0}")'.format(value)
        )
    self._part_of_speech = value

@property
def forms(self):
    return self._forms

@forms.setter
def forms(self, value):
    if not isinstance(value, list):
        raise TypeError(
            '"forms" property must be a list'
        )
    self._forms = value

def __eq__(self, other):
    if (self.lemma == other.lemma
            and self.part_of_speech == other.part_of_speech):
        return True
    return False
```

Формат представления данных

- Внутреннее представление — объект в памяти компьютера (в терминах ООП).

Формат представления данных

- Внешнее представление — любое. Например — XML (хотя вряд ли...).

```
<meta>
    <author>Феодосий Печерский</author>
    <title>Поучение в среду третьей недели поста о терпении и о любви</title>
    <ms>НИОР РГБ. Ф. 256 (Н.П. Румянцев). № 406.</ms>
    <ed>Чаговец 1901</ed>
</meta>

<text>
    <s>
        <br id="103" type="folio" />
        <br id="I" type="page" />
        <w grammar="PRAE" lemma="въ">въ</w>
        <w grammar="N" lemma="срѣда">сре(д)</w>
        <w grammar="ADJ" lemma="3">г҃</w>
        <w grammar="N" lemma="недѣлꙗ">не(д)</w>
        <w grammar="N" lemma="пость">по(с)</w>
        <w grammar="ADJ" lemma="сватыи">ст҃го</w>
```

Вместо лемматизатора

- Славянские языки известны развитой системой флексий и чередований (*могу* — *можешь*). Кроме того, у древнерусского языка нет нормативной орфографии (**ꙁъваша** = **ꙁъваша** = **ꙁваша**…).

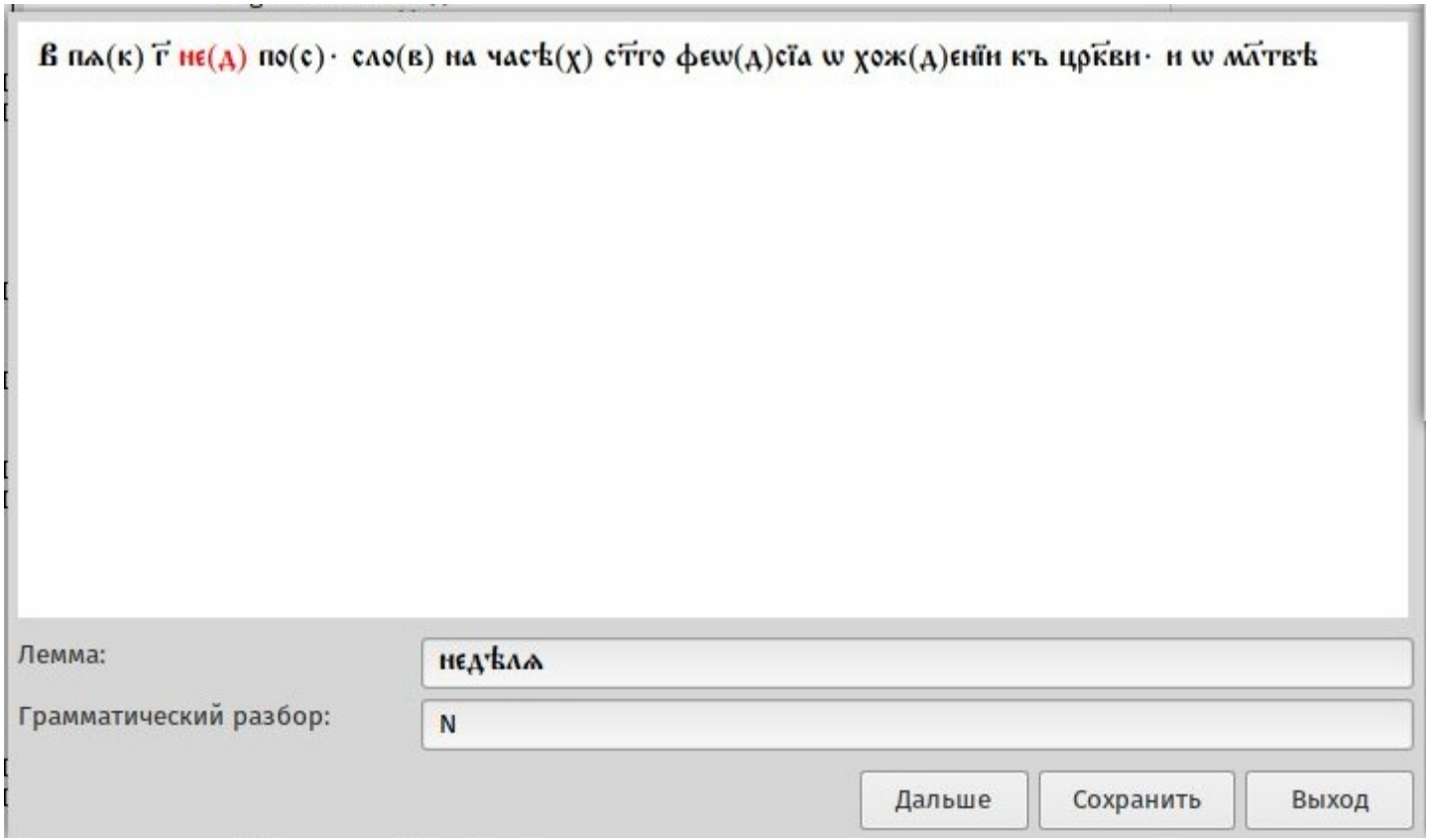
  Расписывать формальные правила разбора таких словоформ — самостоятельная задача, неоправданно большая по трудозатратам.

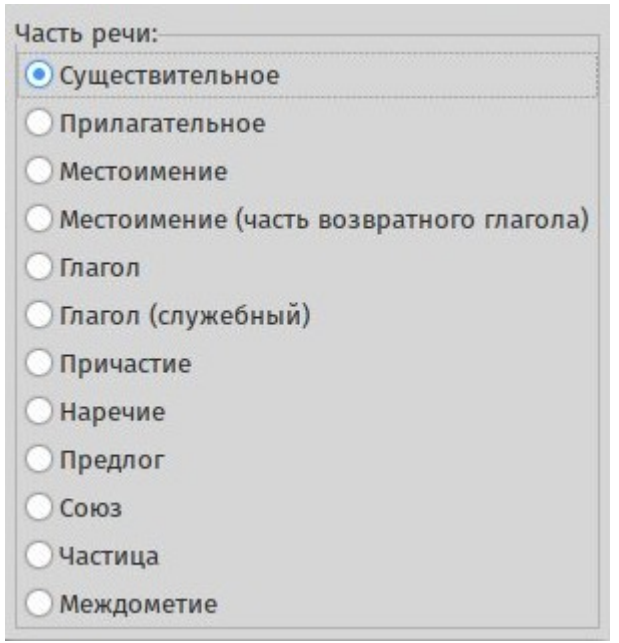
- Однако реализованы

  (1) консольная программа для прецедентного разбора (уже размеченные тексты выступают как референтный корпус).

  (2) GUI для снятия омонимии в визуальном режиме.

Пара скриншотов





Спасибо за внимание!