К проблеме создания размеченных корпусов текстов в графике XIX века

Лебедев Александр Александрович,
Рогов Александр Александрович,
Кулаков Кирилл Александрович,
Москин Николай Дмитриевич
Петрозаводский государственный университет, г. Петрозаводск

Введение

Одна из отличительных особенностей корпуса СМАЛТ (“Статистические методы анализа литературных текстов”) - это его нацеленность на оригинальные публицистические тексты, написанные во второй половине XIX века (60-70 годы), что противопоставляет данный корпус большинству других из числа современных русскоязычных лингвистических корпусов (где тексты ориентированы на современную орфографию).

Подобного рода подход вызывает очевидные затруднения, связанные в первую очередь с проблемами автоматической обработки таких текстов.

Введение

Основу корпуса СМАЛТ составляют тексты статей журнала «Время» (1861-1863 гг.), представленные в дореформенной графике. Часть произведений имеет установленного автора (Ф. М. Достоевский, А. А. Григорьев, М. И. Владиславлев и др.). Авторство других текстов однозначно не определено (эти тексты с пометкой Dubia).

Тексты хранятся в дореформенной графике, но при этом пользователю предлагается и современное их написание. Тексты второй половины XIX века демонстрируют большую орфографическую вариативность, что затрудняет подготовку материала для подобного корпуса (в том числе и в автоматическом режиме).

Пример

В статье «Вопросъ объ университетах» мы можем встретить два разных написания слова «профессор» - через одну и через две «с»:

1) *...по поводу ея **профессоръ** Костомаровъ написалъ свою статью...*

*Корпорація **профессоровъ** сохраняется во всей силѣ…*

2) *...явилось третье мнѣніе г **професора** Стасюлевича…*

*Что обязывало **професора** имѣть въ виду воспитательно-учебныя цѣли*

Многозначное описание

Учет подобных особенностей-разночтений полезен с точки зрения нескольких аспектов: корпус СМАЛТ позволяет учесть изменение орфографических норм, а также проследить за общей логикой развития грамматической системы русского языка XIX века.

Для решения проблемы многозначного описания в программной реализации корпуса СМАЛТ выполнено разделение написания слов и их словоформ. Таким образом, разные написания слов могут иметь одну общую словоформу. Кроме этого, для каждого слова в словоформе приведено современное написание, что позволяет выполнять поисковые запросы без использования дореформенной графики.

Обработка составных словоформ

Один из сложных вопросов, связанных с морфологической разметкой слов в лингвистическом корпусе - это обработка составных словоформ. Рассмотрим несколько примеров, которые могут быть по-разному проанализированы с точки зрения их размещения и обработки в корпусе:

1) *Въ первомъ нумерѣ газеты День* ***помѣщена была*** *статья объ университетахъ покойнаго Хомякова…*

2) *…умноженіе у насъ въ Россіи воспитательныхъ заведеній съ* ***болѣе широкимъ*** *преподаваніемъ чѣмъ въ гимназіи принесетъ намъ пользу…*

3) *… на эту тему можно написать* ***тридцать пять*** *печатныхъ листовъ…*

Пример

В ранних разборах текстов Ф. М. Достоевского данная проблема для приведенных случаев решалась следующим образом: выделенные компоненты разбирались как одно слово, которому приписывались все грамматические категории.

1) *Въ первомъ нумерѣ газеты День* ***помѣщена была*** *статья объ университетахъ покойнаго Хомякова*

Слово: помѣщена была (id=55056)

Начальная форма: ПОМѢЩЕННЫЙ

Часть речи: Причастие

Пример

2) *...умноженіе у насъ въ Россіи воспитательныхъ заведеній съ* ***болѣе широкимъ*** *преподаваніемъ чѣмъ въ гимназіи принесетъ намъ пользу…*

Слово: болѣе широкимъ (id=55101)

Начальная форма: ШИРОКІЙ

Часть речи: Прилагательное

3) *на эту тему можно написать* ***тридцать пять*** *печатныхъ листовъ*

Слово: тридцать пять (id=55392)

Начальная форма: ТРИДЦАТЬ ПЯТЬ

Часть речи: Числительное

Поздние варианты разбора

В поздних вариантах разбора, которые легли в основу корпуса СМАЛТ, такие же контексты разбирались пословно, и у пользователя корпуса есть возможность посмотреть грамматические категории каждого из слов.

На текущий момент для некоторых текстов в корпусе СМАЛТ присутствует морфологическая разметка в двух вариантах (условно называемых «старый» и «новый»), несколько различающаяся по набору частей речи и по грамматическим категориям, однако вне зависимости от выбранного варианта элементы в данных контекстах разбираются как два отдельных слова.

Пример

Рассмотрим разборы со старым набором атрибутов:

1) *Въ первомъ нумерѣ газеты День* ***помѣщена была*** *статья объ университетахъ покойнаго Хомякова*

Слово: помѣщена (id=12210)
Начальная форма: помѣщенъ
Часть речи: Причастие

Слово: была (id=369)
Начальная форма: быть
Часть речи: Глагол
Отвлеченный глагол-связка: Да

Пример

2) *...умноженіе у насъ въ Россіи воспитательныхъ заведеній съ* ***болѣе широкимъ*** *преподаваніемъ чѣмъ въ гимназіи принесетъ намъ пользу…*

Слово: болѣе (id=3763)

Начальная форма: болѣе

Часть речи: Наречие

Вспомогательная часть сложной степени сравнения: Да

Слово: широкимъ (id=12611)

Начальная форма: (более) широкій

Часть речи: Прилагательное

Пример

3) *на эту тему можно написать* ***тридцать пять*** *печатныхъ листовъ*

Слово: тридцать (id=14819)
Начальная форма: тридцать
Часть речи: Числительное
Разряды по значению: Количественное
По способу образования: Часть составного

Слово: пять (id=14820)
Начальная форма: пять
Часть речи: Числительное
Разряды по значению: Количественное
По способу образования: Часть составного

Пример

Рассмотрим разборы с новым набором атрибутов:

1) *Въ первомъ нумерѣ газеты День* ***помѣщена была*** *статья объ университетахъ покойнаго Хомякова*

Слово: помѣщена (id=10571)

Начальная форма: помѣщенный

Часть речи: Причастие

Слово: была (id=10572)

Начальная форма: быть

Часть речи: Глагол

Пример

2) *…умноженіе у насъ въ Россіи воспитательныхъ заведеній съ* ***болѣе широкимъ*** *преподаваніемъ чѣмъ въ гимназіи принесетъ намъ пользу…*

Слово: болѣе (id=5341)

Начальная форма: болѣе

Часть речи: Наречие

Степень сравнения: Компонент сравнительной аналитической степени сравнения

Слово: широкимъ (id=10889)

Начальная форма: широкій

Часть речи: Прилагательное

Форма: Полная

Степень сравнения: Компонент сравнительной аналитической степени сравнения

Пример

3) *на эту тему можно написать* ***тридцать пять*** *печатныхъ листовъ*

Слово: тридцать (id=12656)
Начальная форма: тридцать
Современное написание: тридцать
Часть речи: Числительное

Слово: пять (id=1116)
Начальная форма: пять
Современное написание: пять
Часть речи: Числительное

Составные союзы

Второй подход, при котором каждая подобная словоформа разбирается по отдельности, видится более удобным с точки зрения хранения и поиска данных. В то же время, разумным представляется указание в ходе разбора на то, что некоторое слово является, к примеру, частью составного союза:

мы и сами не предлагаемъ никакихъ реформъ ***да и*** *о предложенныхъ реформахъ не очень распространимся*

Пример

Слово: да (id=2179)
Начальная форма: да
Часть речи: Союз
По составу: Часть составного союза

Слово: и (id=2180)
Начальная форма: и
Часть речи: Союз
По составу: Часть составного союза

Заключение

Каждый из представленных подходов имеет свои достоинства и недостатки и не может быть выбран в качестве основного. Таким образом, разумным решением является предоставление гибких инструментов анализа текстов.

Программная реализация корпуса СМАЛТ позволяет выполнять морфологическую разметку текста в удобном для исследователя формате. Также планируется внедрить возможность использования авторской морфологической разметки и последующего сопоставления морфологических разборов.

Спасибо за внимание!

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-012-90026.