Описание лингвистического компьютерного ресурса.

Выполнили:
Полянская Анна (БКЛ 171) *off.polyanskaya.a@gmail.com*
Стратулат Екатерина (БКЛ 172) *stratulat.ekaterina1999@gmail.com*
Щербакова Анна (БКЛ 172) *aniezka.sherbakova@gmail.com*

ресурс: Тестовый корпус с параллельной синтаксической разметкой

Введение

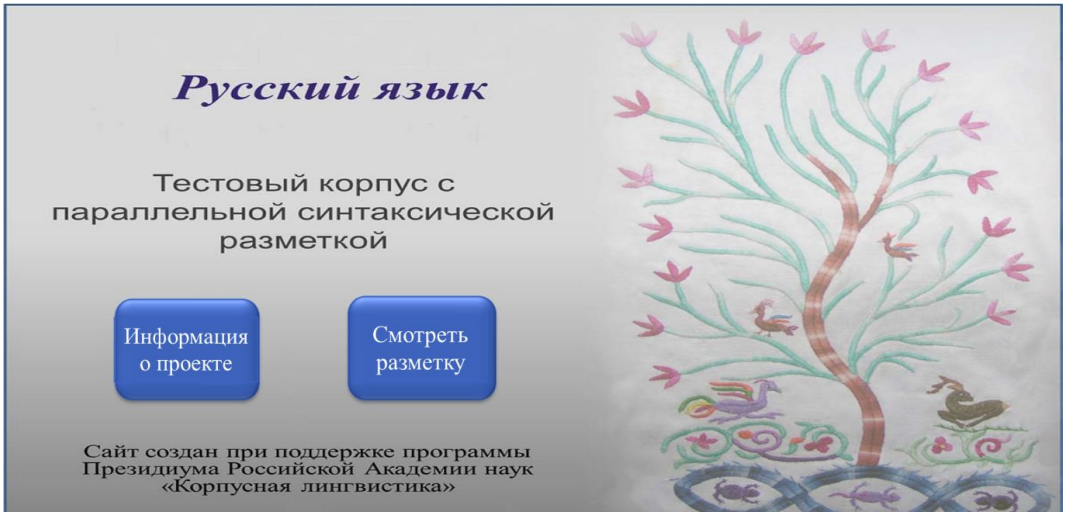
Для экспертизы мы выбрали Тестовый корпус с параллельной синтаксической разметкой (RUSSIAN SYNTAX TREE BANK). RSTB – банк синтаксических деревьев. В нем представлены результаты разбора 64800 предложений (1 млн словоупотреблений) тремя автоматическими системами синтаксического анализа: SyntAtom, SemSin, Russian Malt. В корпус вошли предложения из текстов разных жанров, включая научную и художественную литературу, а также тексты новостных сообщений.

На сайте также представлено 800 предложений из этого корпуса, выбранных случайным образом и размеченных вручную.

Все три системы используют синтаксическое представление в виде деревьев зависимостей. Узлами дерева являются слова предложения. Направленными стрелками соединены два слова, находящиеся в отношении синтаксической связи. Направление стрелок – от главного к зависимому. Сохранены исходные названия синтаксических связей и морфологические пометы словоформ, используемые в соответствующих системах.

Ресурс создавался при поддержке программы Академии наук РФ «Корпусная лингвистика»

Дизайн



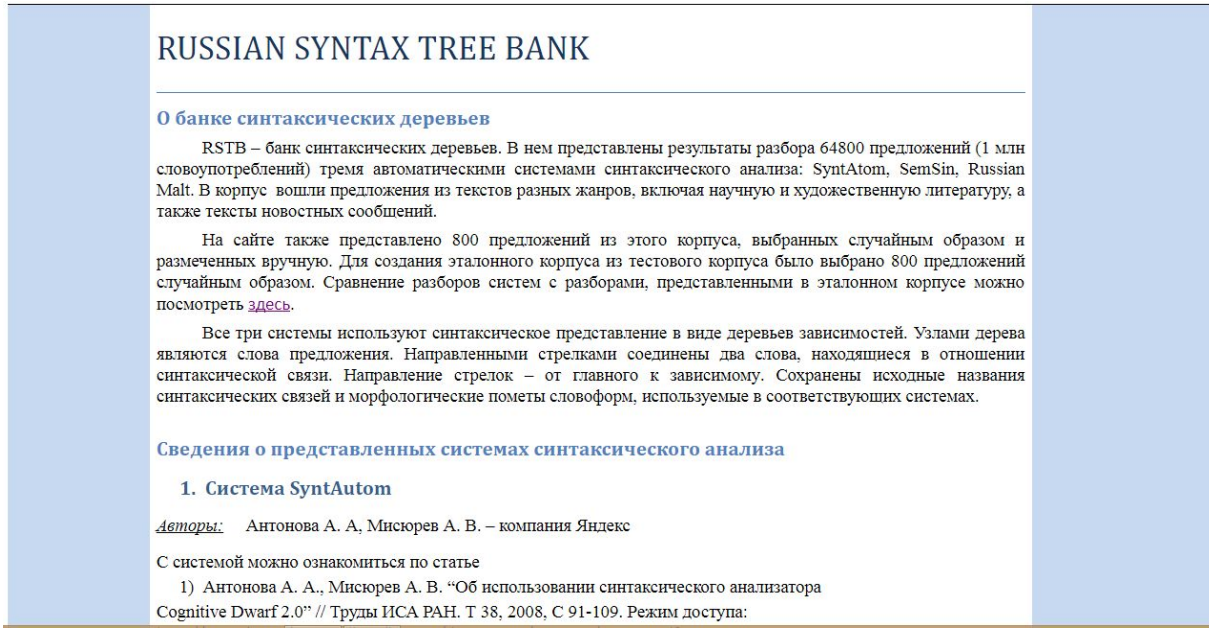
Главная страница выполнена в приятной цветовой гамме, дизайн достаточно простой, картинка в правой части отсылает к понятию синтаксических деревьев. На главной странице присутствуют 2 кнопки: “Информация о проекте” и “Смотреть разметку”. Кнопки - просто картинки с ссылками, курсор не меняется при наведении на них. Страница не масштабируется. Сайт удобен в использовании с планшета/ смартфона.

Нажимаем “Смотреть разметку”. Перед нами

На данной страницы всё довольно просто: белый фон, вверху расположен заголовок “Tree Bank” . Использованы дефолтные кнопки, разделители, таблицы.
Слева расположена колонка с системами синтаксического анализа, по которым производится поиск.
По центру расположена поисковая строка.

Ресурс глазами новичка и помощь пользователю

На главной странице сайта есть кнопка “Информация о проекте”. При переходе на этот раздел мы видим следующее:



RUSSIAN SYNTAX TREE BANK

О банке синтаксических деревьев

RSTB – банк синтаксических деревьев. В нем представлены результаты разбора 64800 предложений (1 млн словоупотреблений) тремя автоматическими системами синтаксического анализа: SyntAtom, SemSin, Russian Malt. В корпус вошли предложения из текстов разных жанров, включая научную и художественную литературу, а также тексты новостных сообщений.

На сайте также представлено 800 предложений из этого корпуса, выбранных случайным образом и размеченных вручную. Для создания эталонного корпуса из тестового корпуса было выбрано 800 предложений случайным образом. Сравнение разборов систем с разборами, представленными в эталонном корпусе можно посмотреть здесь.

Все три системы используют синтаксическое представление в виде деревьев зависимостей. Узлами дерева являются слова предложения. Направленными стрелками соединены два слова, находящиеся в отношении синтаксической связи. Направление стрелок – от главного к зависимому. Сохранены исходные названия синтаксических связей и морфологические пометы словоформ, используемые в соответствующих системах.

Сведения о представленных системах синтаксического анализа

1. Система SyntAutom

Авторы: Антонова А. А, Мисюрев А. В. – компания Яндекс

С системой можно ознакомиться по статье

1) Антонова А. А., Мисюрев А. В. “Об использовании синтаксического анализатора Cognitive Dwarf 2.0” // Труды ИСА РАН. Т 38, 2008, С 91-109. Режим доступа:

Здесь приводится описание банка синтаксических деревьев и даются сведения о представленных системах синтаксического анализа. Далее сказано о ручной разметке, в которой присутствуют ссылки с дополнительной полезной информацией.

О ручной разметке

Ручная разметка 800 предложений производилась двумя независимыми аннотаторами в соответствии с инструкцией по ручной разметке. Инструкция была разработана Е. Г. Соколовой в рамках курса по Автоматической обработке текста, читаемого в Институте лингвистики РГГУ. С последней версией инструкции можно ознакомиться здесь.

Например, нажимаем на “здесь”. Перед нами открывается окно с вариантами разметки синтаксических конструкций:

Варианты разметки синтаксических конструкций

Сопоставительный анализ

Е.Г.Соколова

Институт лингвистики, РГГУ

Разметка основных конструкций, в которых возникают варианты установления и направления связей

P1, P2 … - варианты разбора разными системами

Напр. Связи	пример	Эталон	P1	P2	P3	P4	P5	P6
1. Prep+S	На горе	→	→	→	←	→	→	←
2. Nk+S	десять минут два урока	→ →	Нет прим	→	←. ←	← →	→	← ←
Nk1+Nk2+Nk3+S	Сто тридцать шесть человек	Nk1←Nk2←Nk3 Nk3→S						
3. Nk+Np+S	В шестьдесят первом году	←←		← ←	←←	ОШ←	→→	←←
4. Дата: N+год	(в) 1774 году	→		←	→	←	→	←
5. Cc+M2	и M2	←	→	←сост!	-	←	←	←
6. M1+Cc	(M1 и)	-	←	-	→	-	-	-
7. G+M1+Cc+M2	Увидел M1 и M2	G→M1 M1→M2	G→Cc	G →M1 M1→ M2	G →M1 M1→ M2	G → M1 M1→ M2	G→M1 G→M2	G→M1 G→M2
8. Vбыть+КрПрич	Было создано	←	→	←сост!	←	→	→	←
9. (-быть)+Инф	Стали нападать	→	→	→	←	→	→	→
10. Вопрос с вопр. местоим: Q+V?	Почему стали нападать?	←	←	??	←	→	←	←

Приводится расшифровка сокращений:

Существительное с предлогом (1):
S – существительное
Prep – предлог
-
Количественные конструкции (2-4):
N – числительное, представленное числом;
Nk – количественное числительное, представленное словом;
Np – порядковое числительное, представленное словом;
-

Сочинительная конструкция (5-7)
G – хозяин сочиненной цепочки
Cc – сочинительный союз
M1 – первый членили член, предшествующий сочиненному
M2 – сочиненный член
Cs - Подчинительный союз

V1 – сказуемое главного предложения

Это очень помогает новичкам, но найти данные подсказки было не очень просто. Выгоднее расположить кнопку “Помощь” на главной странице сайта.

В сравнении с НКРЯ в данном корпусе очень мало подсказок. Не помешали бы скриншоты с примерами по поиску в корпусе и какие-либо инструкции или пояснения на самой странице поиска.

Расшифровки всех терминов и сокращений находятся на внешних страницах, не на все из них есть ссылки, например, система Этап-3 вообще не описывается на странице с информацией, а расшифровку её терминов приходится смотреть в НКРЯ.

Функционал

Доступ к корпусу бесплатный и не требует регистрации, оффлайн доступ не предусмотрен. Сайт находится в домене ОТиПЛа МГУ. При запросах с большим количеством результатов поиск работает достаточно медленно.

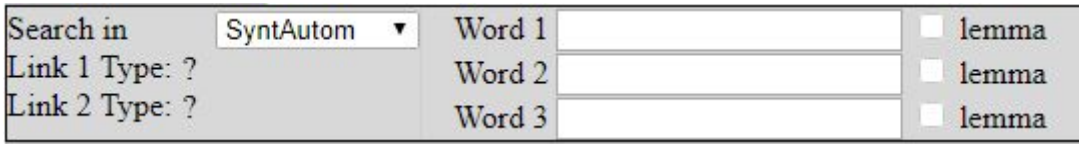
Слои разметки:

- морфосинтаксическая разметка (помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана его синтаксическая структура)
- лемматизация (приведение слова к начальной форме)
- токенизация (разбиение текста на единицы (слова, знаки препинания) для дальнейшей обработки)

1. Поиск

Если выбрать “*Numbers*”**,** то поиск предложений будет производиться по их порядковым номерам. Мы можем выбрать нужные нам предложения из 64853. *“From sentence”* выбирает с какого предложения начинается поиск. “*Quantity”* выбирает количество предложений.

Переключив на “*Search*”**,** мы видим следующее



Здесь можно выбрать в какой системе искать, можно задать 3 слова для поиска, а также указать тип синтаксической связи и морфологические признаки слов (нажать на “*lemma”*).

Возможен поиск с учётом словоформ, однако отсутствует возможность поиска части слова с помощью операторов (например, “мам*”). Отсутствует какая-либо информация о предложениях (например: источник, автор, дата, жанр текста и т.п.).

Есть поиск по нескольким леммам сразу, однако нет возможности задавать желаемое расстояние между словами. Нет подкорпусов.

Примеры поисковых запросов:

1. word1: fem , nom , nn
   word2: nn

   потом link1 type: homo

2. word1: nn , fem , nom
   word2: adj , fem , nom

   потом word1: поездка

3. word1: нельзя
   word2: будет

   потом link1 type: inf

2. Результаты и выдача:

Результаты поискового запроса отображаются списком. Для каждого предложения указан его порядковый номер в общем банке деревьев.

Существует ограничение по количеству примеров в выдаче - до 100 предложений. При вводе числа больше 101 в поле “Количество” система выдаёт ошибку.

Существенным минусом является отсутствие возможности скачать полученные результаты в каком-либо виде, поэтому достаточно сложно использовать примеры из корпуса в большом количестве.

Результаты выдачи нельзя никак упорядочить или сортировать, нет возможности пользоваться методом KWIC (key word in context).

В выдаче искомые слова выделяются в рамку.
При наведении на любое слово из предложения мы можем видеть его синтаксические и морфологические признаки (в виде подсказки). Главное слово в паре подсвечивается желтым цветом, зависимое слово выделяется красным цветом.

Можно визуализировать результаты поиска. Синтаксическая структура предложения представляет из себя схему зависимостей и отображается в виде стрелочек над текстом. Каждая такая стрелка идет из главного слова к зависимому. Отношения связывают только отдельные слова, а не словосочетания.

К каждому предложению можно открыть таблицу, в которой указаны типы семантической связи между всеми токенами.

Заключение

Данный ресурс, обеспечивающий сравнение разметки нескольких анализаторов, позволит выявить наиболее надежно устанавливаемые синтаксические связи, а также наиболее сложные и проблемные места синтаксической разметки для русского языка независимо от реализуемого в системе подхода. На нем можно отработать технологию синтаксического поиска для корпуса большого объема, выявить наиболее актуальные потребности пользователей, что касается синтаксических запросов. Корпус позволит апробировать методы автоматического или полуавтоматического исправления ошибок автоматических разметчиков.