

СОСТАВ И СТРУКТУРА НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Яхья-заде А.С.

*Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь*

Петрова Н.Е. – к.филол.н., доцент

Рассматривается национальный корпус русского языка как уникальный лингвистический ресурс, предоставляющий доступ к обширной коллекции текстов на русском языке. Он включает в себя не только современные и исторические тексты, но также и многочисленные диалекты, представленные в различных стилистических и функциональных регистрах.

Национальный корпус русского языка (НКРЯ) – это собрание независимых корпусов, каждый из которых предназначен для решения определённых лингвистических задач. Каждая из этих коллекций текстов является большой по объёму и представительной, что делает их ценным материалом для количественных и качественных исследований [1].

Наиболее общим по своим задачам является основной корпус [2], включающий прозаические письменные тексты на протяжении последних трёх с лишним веков. В основной корпус входят и опубликованные, и рукописные, и электронные тексты самых разных жанров. Разметка основного корпуса носит наиболее нейтральный характер, как и отдельного газетного корпуса [3], самого большого в НКРЯ. Газетный корпус включает тексты центральных СМИ начиная с 1980-х годов.

Все тексты, входящие в основной корпус, содержат метаразметку, морфологическую, словообразовательную, синтаксическую и семантическую разметку.

Метаразметка основного корпуса включает сведения о названии текста, дате его создания, имени, годе рождения, поле автора, месте и дате публикации, источнике, по которому дается текст, его сфере функционирования, жанре и типе текста, хронотопе художественных произведений и мемуаров, специфике аудитории, орфографии и о типе морфологической разметки. Возможно отобрать подкорпус по всем этим параметрам.

Морфологическая разметка для русского языка осуществляется с помощью специальных программ автоматического морфологического анализа и лемматизации. Большинство текстов размечены параллельно двумя программами, MyStem и Rubic, специально адаптированными для обработки текстов разных сфер, жанров и времени создания [4].

Русскоязычная модель MyStem основана на электронном грамматическом словаре и умеет строить гипотезы для несловарных слов. Ряд словоформ, часто встречающихся в текстах корпуса, но не предусмотренных этим анализатором, получает индивидуальный разбор, заданный списком или построенный по правилам. Каждой словоформе приписываются все предусмотренные словарем или правилами разборы, независимо от контекста – поиск по разметке MyStem доступен в режиме «Все разборы».

Нейросетевая модель Rubic обучена на представительном наборе текстов, размеченных вручную, и предлагает для словоформы один, наиболее вероятный в данном контексте, набор грамматических признаков. Rubic также автоматически строит правила преобразования словоформы в лемму, а затем из нескольких наиболее вероятных в данном контексте гипотез отбирается та, которая входит в составленный экспертами словарь. Если такого совпадения нет, то словоформе приписывается наиболее вероятная с точки зрения модели лемма. Дополнительными правилами исправляются ошибочные разборы для сочетания леммы и части речи, встретившиеся в корпусе более 40 раз. Это, в частности, позволяет дать корректные разборы для частотных архаичных, просторечных и орфографически искаженных форм.

Нейросетевая модель Rubic осуществляет также синтаксическую разметку основного корпуса, предсказывая один синтаксический разбор для всего предложения. Анализатор строит для предложения дерево зависимостей, в котором каждое слово связано дугой со своим синтаксическим «хозяином», за исключением вершины дерева – главного слова предложения. Дуги имеют метку синтаксического отношения между главным и зависимым словом. На основе этого дерева дополнительные правила определяют в предложении синтаксические составляющие — клаузы и группы. В настоящий момент в корпусе доступен поиск по синтаксической роли слова как зависимого и основным типам составляющих, а в дальнейшем планируется подключить полный поиск по дереву зависимостей.

В то же время поэтический корпус [5] служит основой для специальных стиховедческих исследований, поэтому в нем есть и особая разметка, связанная с ключевыми для стиховедения понятиями — метром и ритмом. Акцентологический корпус [6] посвящен истории русского ударения, играющего ключевую роль в его разметке. Аннотация устного корпуса [7] также отражает ударения и иную специфику звучащей речи. Промежуточное положение между устным и письменным корпусом занимает корпус социальных сетей [8], тексты которого меньше связаны ограничениями литературной нормы, а также активно используют такую специфическую знаковую систему, как эмодзи. В мультимедийном корпусе [9] текст сопровождается синхронизированной видео- или аудиозаписью, а в нескольких кинофильмах размечены и жесты. В синтаксическом корпусе [10] предложениям сопоставлена специальная сложная разметка синтаксической структуры. Обучающий корпус [11] включает в себя жанровую и морфологическую разметку, адаптированную к российской школьной программе. Корпус «От 2 до 15» [12], посвященный детскому и подростковому чтению, включает автоматически размеченную информацию о возрасте читателей каждого текста.

Помимо текстов на современном русском языке, ориентированном на литературный стандарт, НКРЯ стремится представить русский язык в его историческом и географическом многообразии. Так, целый ряд корпусов – исторические [13]. Отдельными коллекциями текстов представлен древнерусский язык, старорусский и церковнославянский язык в его русской версии. Исторические тексты и основной корпус связаны общим поиском – это так называемый панхронический корпус [14], который позволяет проследить историю слова или грамматической конструкции на протяжении нескольких веков.

Диалектный корпус [15] включает в себя устные тексты, записанные у носителей традиционных русских диалектов по всей России, в фонетической записи и с сохранением всех особенностей лексики и грамматики.

Большинство корпусов, входящих в НКРЯ, одноязычные, то есть в них входят только тексты на одном языке. Исключением является параллельный корпус [16], где оригинальные русские тексты сопровождаются переводом на другой язык или иноязычные произведения переведены на русский. НКРЯ включает несколько десятков русско-иноязычных языковых пар и многоязычный корпус, где один и тот же текст переведен на несколько языков. Один из исторических корпусов — корпус берестяных грамот – тоже параллельный: древнерусский текст дается с переводами на современный русский и английский языки. Наконец, имеется и мультимедийный параллельный корпус, где представлены или англоязычные фильмы в русском переводе, или постановки одного и того же текста на английском и русском языках.

Таким образом, НКРЯ является незаменимым инструментом для изучения изменений в русском языке, анализа грамматических и словообразовательных структур, а также для проведения сравнительного анализа между русским и другими языками через параллельные тексты. Этот ресурс обладает значительным потенциалом для специалистов в области лингвистики, филологии и смежных дисциплин, а также может быть использован в образовательных целях.

Список использованных источников:

1. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/>. – Дата доступа: 04.05.2024.
2. Основной корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/main>. – Дата доступа: 04.05.2024.
3. Газетный корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/paper>. – Дата доступа: 04.05.2024.
4. Программное обеспечение НКРЯ [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/page/corpora-about/>. – Дата доступа: 07.05.2024.
5. Поэтический корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/poetic>. – Дата доступа: 07.05.2024.
6. Акцентологический корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/accent>. – Дата доступа: 07.05.2024.
7. Корпус устной речи [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/spoken>. – Дата доступа: 07.05.2024.

8. Корпус социальных сетей [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/search?search=CgQyAggWMAE%3D>. – Дата доступа: 07.05.2024.
9. Мультимедийный корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/murco>. – Дата доступа: 07.05.2024.
10. Синтаксический корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/syntax>. – Дата доступа: 08.05.2024.
11. Обучающий корпус русского языка [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/school>. – Дата доступа: 08.05.2024.
12. Корпус «От 2 до 15» [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/kids>. – Дата доступа: 08.05.2024.
13. Исторический корпус [Электронный ресурс]. – Режим доступа: https://ruscorpora.ru/corpus/old_rus. – Дата доступа: 08.05.2024.
14. Панхронический корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/panchron>. – Дата доступа: 10.05.2024.
15. Диалектный корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/dialect>. – Дата доступа: 10.05.2024.
16. Параллельный корпус [Электронный ресурс]. – Режим доступа: <https://ruscorpora.ru/corpus/para-eng>. – Дата доступа: 10.05.2024.