

Что такое Национальный корпус русского языка?

Корпус – это собрание текстов в электронной форме, в котором можно осуществлять поиск слов, словосочетаний, грамматических форм, значений слов с помощью определенной поисковой системы. Существует много разновидностей корпусов – например, корпус одного автора, корпус одной книги (в частности, первые корпуса делались для Библии). **Национальный корпус** некоторого языка представляет данный язык на определенном этапе его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т. п. Национальный корпус русского языка функционирует в Интернете с 2003 года и в данный момент включает самые разные русские тексты общим объемом 140 миллионов словоупотреблений. В перспективе Национальный корпус русского языка должен содержать тексты общим объемом порядка 200 миллионов словоупотреблений.

Национальный корпус русского языка, как и национальные корпуса других языков, имеет две важные особенности. Во-первых, он **представителен и сбалансирован**, т. е. содержит все типы письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую **разметку**, или **аннотацию**). Разметка — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном Интернете. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. В Национальном корпусе русского языка в настоящее время широко используются три типа разметки: метатекстовая (она характеризует текст в целом, с точки зрения его автора, жанровых и других особенностей), морфологическая и семантическая (морфологическая и семантическая разметки характеризуют не целый текст, а отдельное слово).

Зачем нужен национальный корпус?

Национальный корпус предназначен в первую очередь для научных исследований лексики и грамматики языка. Другая задача корпуса — предоставление всевозможных справок, относящихся к указанным

областям (лексика, грамматика, акцентология, история языка). Раньше специалист мог лишь просматривать тексты и вручную выписывать из них нужные примеры; эта деятельность была очень трудоемкой и не позволяла обрабатывать большие массивы материала. Теперь ограничений на объем анализируемого материала и скорость поиска информации нет, а это означает, что в распоряжении ученого и преподавателя оказываются колоссальные массивы текстов самого разного типа.

Основными пользователями национальных корпусов являются, конечно, исследователи-лингвисты самого разного профиля. Однако ими круг пользователей корпуса вовсе не ограничивается! Надежные статистические данные о языке определенной эпохи или определенного автора могут интересовать литературоведов, историков и представителей многих других областей гуманитарного знания. Важнейшее значение национальные корпуса имеют и для преподавания языка в качестве родного или иностранного; всё больше учебников и учебных программ в наше время оказываются ориентированы на корпус. Быстро и эффективно проверить с помощью корпуса особенности употребления незнакомого слова или грамматической формы у авторитетных авторов сможет и иностранец, и школьник, и учитель, и журналист, и писатель.

Как будет развиваться Национальный корпус?

Национальный корпус русского языка будет охватывать прежде всего период от начала XIX до начала XXI века: этот период представляет язык в разных социолингвистических вариантах — литературном, разговорном, просторечном, отчасти диалектном. В корпус включаются оригинальные (непереводные) произведения художественной литературы (проза и драматургия, поэзия), имеющие культурную значимость, а также представляющие интерес с точки зрения языка. Но Национальный корпус ни в коей мере не является только корпусом языка художественной литературы. Помимо художественных текстов, в корпус в большом количестве включаются и другие образцы письменного (а для современного этапа — и устного) языка: мемуары, эссеистика, публицистика, научно-популярная и научная литература, публичные выступления, частная переписка, дневники, документы и т. п.