Методы автоматического анализа естественного языка – это одни из наиболее развивающихся и актуальных областей компьютерной лингвистике. В то же время и здесь остаются задачи, для которых не найдено единого решения, в том числе морфологический и синтаксический анализ текстов на русском языке. Одним из таких решений является гибридный анализатор NLTK4RUSSIAN на основе NLTK и PyMorphy2. Это лингвистический комплекс для исследования русскоязычных корпусов текстов, разрабатывающийся на кафедре математической лингвистики СПбГУ. Инструменты комплекса позволяют выполнять главные этапы цикла автоматической обработки текстов – морфологический и синтаксический анализ.

В докладе описывается эксперимент по тестированию гибридного морфологического морфоанализатора на данных соревнований Dialogue Evaluation 2017. Качество анализа оценивается с помощью метрики точность средства измерений (accuracy).

Основные задачи эксперимента: установка и исследование основных режимов работы морфоанализатора NLTK4Russian; создание конвертера, преобразующего теги, используемые для морфологической разметки в Universal Dependencies, в теги формата PyMorphy; обучение морфоанализатора и его тестирование на выбранных данных; обработка результатов и оценка качества общей морфологической и отдельно частеречной разметок с помощью метрики точность средства измерений (accuracy); сравнение полученных данных для использованных корпусов.

Для обучения морфоанализатора были взяты материалы Национального корпуса русского языка и проекта «Открытый корпус». Для тестирования были использованы тексты, которые являлись тестовыми для участников соревнований Dialogue Evaluation 2017, а именно корпусы текстов новостного сайта Lenta, социальной сети ВКонтакте и тексты автора Олега Зайончковского (JZ).

В ходе исследования были получены следующие результаты: точность морфологической разметки практически не зависит от выбранного обучающего корпуса (например, результат частеречной разметки корпуса текстов автора JZ с Национальным корпусом русского языка в качестве обучающего – 74,7%, с корпусом проекта «Открытый корпус» – 74,8%). При этом отмечается тот факт, что точность частеречной разметки в среднем оказалась выше, чем точность полной морфологической разметки; для частеречной разметки средний показатель равен 90,3%, а для полной морфологической разметки среднее значение заметно меньше – 69,7%. Наконец результаты полной морфологической разметки варьируются в зависимости от жанра текстов: наибольшей точности удалось добиться на текстах автора JZ (74%), следом идёт корпус текстов социальной сети ВКонтакте (69%); наименьшая точность наблюдается для текстов новостного корпуса (65%).