Возможные задачи:

- грамматический разбор (часть речи и тд)

- нормализация (лемматизация)

- синтаксический парсинг

1) Голосующая система (выбирает один из результатов существующих систем)

2) Создать систему с нуля из готовых библиотек (например, лемматизатор, тэггер)

3) Эксперименты с данными на входе

Представления слов в виде символьных эмбеддингов, определение близких слов

Оценить размер данных

Читать: результаты соревнований conll 2017, 2018, статьи про разные системы

Проблемы: пунктуация, смешивание частей речи

confusion matrix (какие части речи систематически принимаются за другие?)

DET – притяжательные прилагательные, указательные местоимения и тд

Пока UDpipe

корпус 10 млн

Обучить юдипайп на этом корпусе, потестировать, сделать confusion matrix для частей речи (есть специальные библиотеки в питоне)

<http://wiki.apertium.org/wiki/UDPipe> – обучение юдипайп

Потом – другие системы