Пост состоит из двух частей. В первой части автор предлагает обратиться к известным статьям, посвященным автоматическому анализу естественного языка, написанным в конце прошлого века (“old school NLP). Несмотря на то, что современные методы анализа сильно отличаются от тех, которые предлагались в 1980-х годах, автор находит такое чтение полезным: основной задачей в тот период являлась задача репрезентации (и фактически формализации) знания. Так в статье Манн и Томпсон рассматриваются модели дискурсивных отношений между клаузами, а в статье Венди Леннерт представлена структура сюжета нарратива. Эти модели, являются на взгляд автора, более сложными, чем те, которые используются для машинного обучения. Более того, автор считает полезным вначале использовать методы, формализующие наши знания, а лишь затем реализовывать эти формализмы с помощью машинного обучения.

С моей точки зрения, рассматриваемая проблема интересна, во-первых, тем, как связаны разные периоды достаточно молодой и бурно развивающейся области компьютерной лингвистики. Интересно, что обращение к казалось бы устаревшим моделям может обогатить современные подходы. Во-вторых, чрезвычайно важной мне показалась критика современных работ, которые выполняются без предшествующего моделирования: данные сначала обсчитываются и лишь потом производится попытка интерпретации результатов.

Во второй части автор рассказывает про свой эксперимент связанный со сравнением длительности произнесения слов на разных языках. Гипотеза состоит в том, что такая оценка может улучшить модели синхронного перевода. Автор использует словари и эвристики, которые уточняют длину слова, а также готовые модули синтеза речи, с помощью которых он выводит зависимость между признаками слова и времени его произнесения. В результате, выводится формула примерно оценивающая время, необходимое для произнесения предложения.

С моей точки зрения, при разметки моделей можно было бы учитывать частотность слов, а также стиль речи. В частности, существует ряд разговорных выражений, которые произносятся быстрее.