Инструкция по разметке корпуса:

Разметка корпуса делится на 6 категорий, некоторые категории могут быть вложенными друг в друга (например, сленговые слова могут встречаться внутри синтаксических конструкций).

Имена героев, полные и неполные (full_name, short_name)
В категорию неполных имён входят все вариации имён героев, не совпадающие с полными (например, Женёк, Вова). При этом упоминания героев только по фамилии (Онегин, Ленский) не входят в эту категорию и не считаются сленговыми.

Сленговые слова (slang)
Сленговыми словами считаются все несловарные, разговорные слова, а также видоизменённые словарные слова (чё). Нецензурные слова и близкие к ним также входят в эту категорию. К сленговым словам по возможности в разметке добавляются литературные аналоги.

Синтаксические конструкции (syntax)
В категорию синтаксических конструкций входят конструкции любого типа с использованием сленговых слов (по приколу) и без них, но имеющих разговорный стиль (по факту, он такой говорит ему).

Цитаты из текста (quote)
Цитатами считаются прямые цитаты из романа (не в шутку занемог), прямые цитаты с перепутанным порядком (я ему вовек буду верна) и цитаты с ошибками (они сошлись, огонь и пламя) с исправлениями, если есть.

Слова-паразиты (дискурсивные маркеры, discourse markers)
В слова-паразиты входят все вводные несловарные обороты (короче, как бы, вот, ну) и сравнительные слова (типа).

Непонятно, что, но интересно (interesting)
Эта категория самая размытая, пока что разметка для неё приветствуется, но необязательна. Сюда входят цитаты из фикшна (например, тикток), некорректные варианты словарных слов (дворянины, успокоенный), непонятный контекст использования слова.

Мета-информация
Класс и профиль обучения говорящего.