Формулы читабельности:

1. LIX = (количество слов в тексте / количество предложений в тексте) + (количество слов длиннее 6 букв \* 100 / количество слов)  ***показала наибольшую корреляцию***
2. Возможно добавить еще одну для сравнения, ***добавлю до вторника, 19 февраля***

Лексика:

1. Максимальная длина слова в морфемах morph/N ***РКИ делает список сложных морфем, за них давать штрафы***
2. Отношение количество типов слов к общему количеству токенов в тексте - Type-Token Ratio (TTR) = word types T / number of tokens N ***лексическое разнообразие***
3. Пропорция абстрактных слов (регулярками по суффиксам) N abstract/N
4. Среднее количество значений на токен (Avg. nr. of senses per token - использовали SALDO — шведская альтернатива Princeton WordNet)

Морфология:

1. Разнообразие лиц - чем больше, тем сложнее
2. Отношение нефинитных глаголов, инфинитивов, бы, себя, причастия, деепричастия ко всем глаголам
3. Количество бы, не, ни (проблема отрицания)
4. Количество это, референтных слов ***кореференция - это сложно***

Синтаксис:

1. Средняя длина зависимостей (Maltparser)
2. Количество именных групп на предложение
3. Количество глагольных групп на предложение
4. Количество предложных групп на предложение
5. Нестандартный порядок слов: например, подлежащее после сказуемого ***РКИ делает список часто встречаемых порядков слов***

***Итого: 15 показателей***