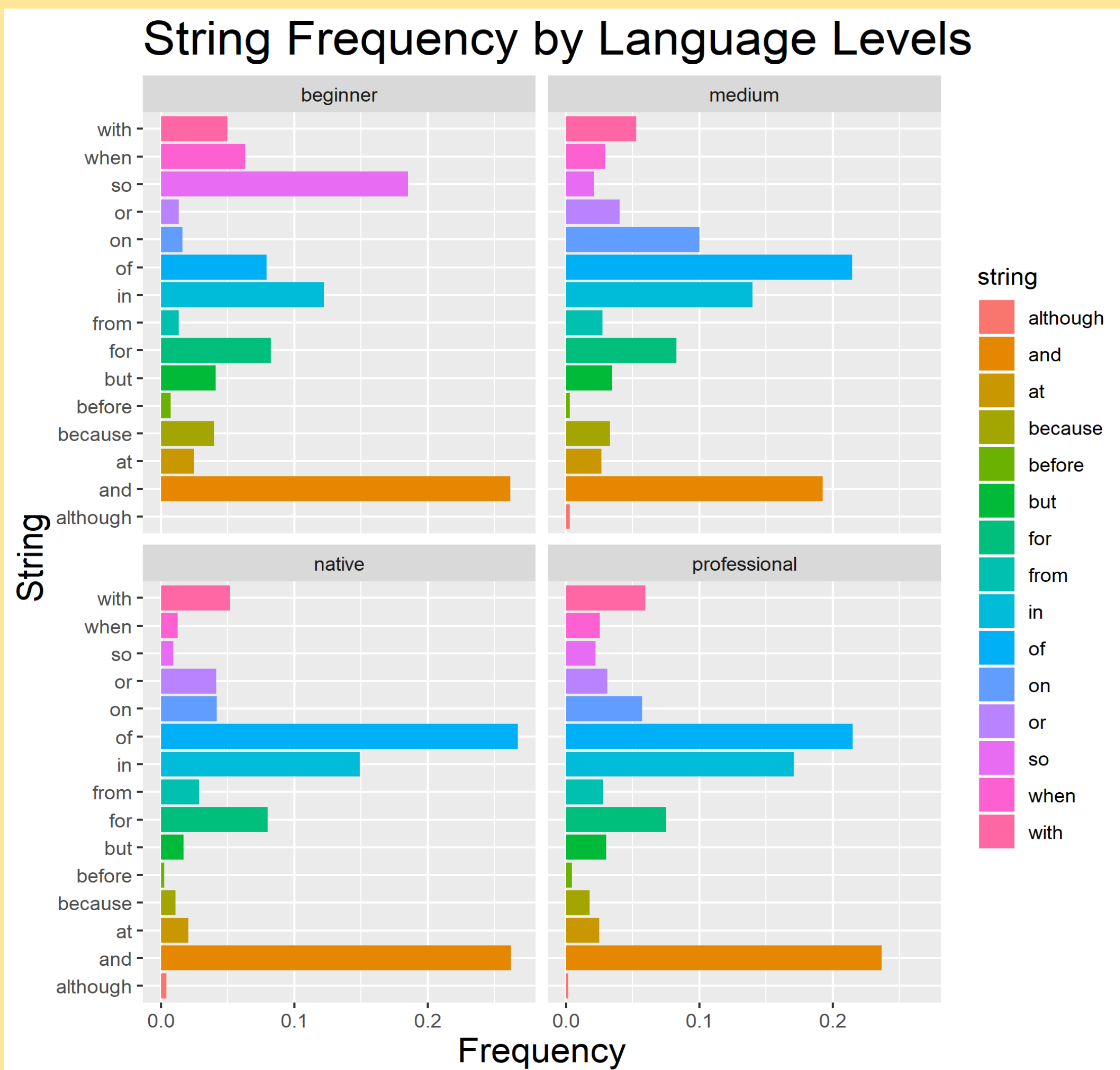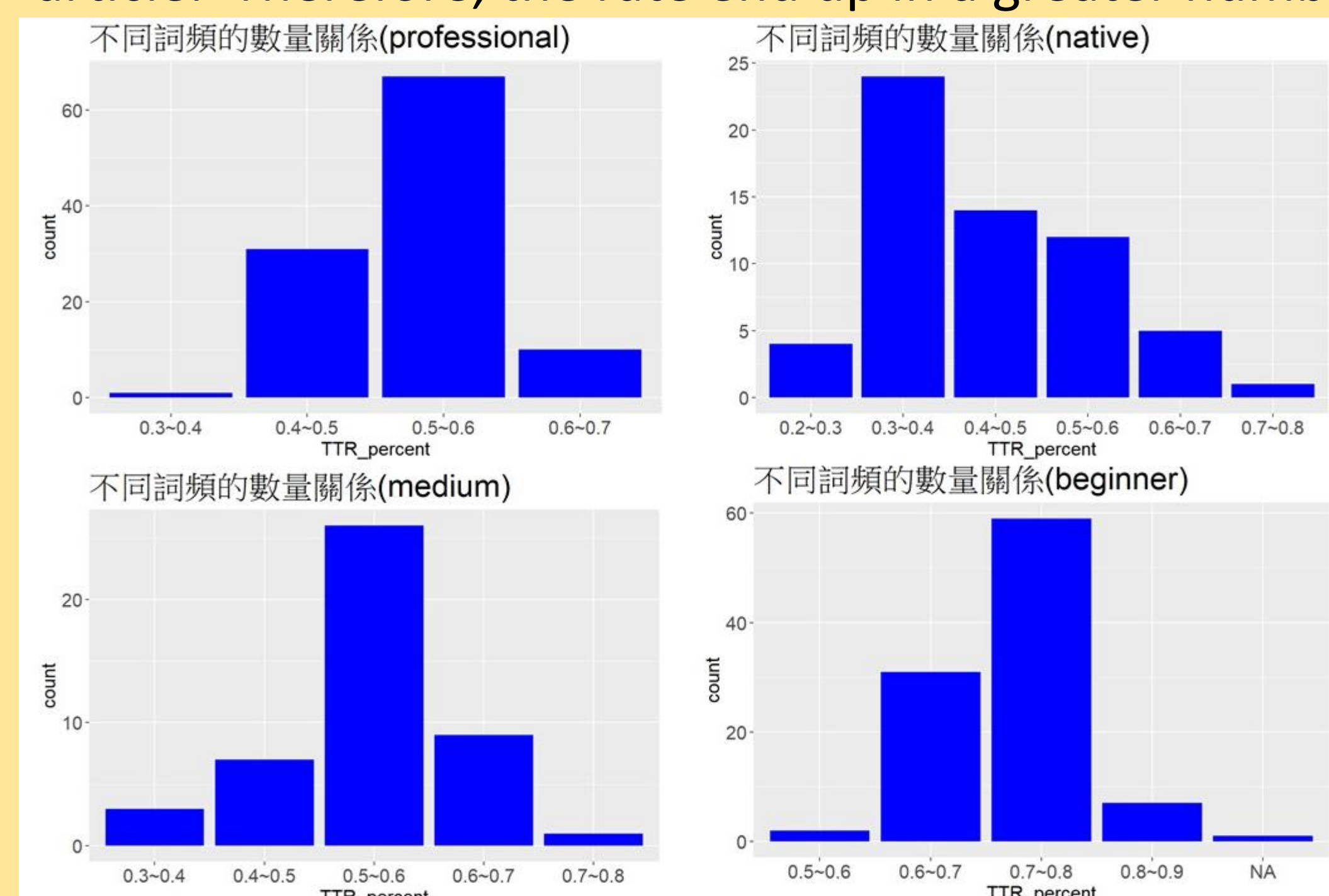# EN Analysis & Grader

Frequency of prep. and conj.

Calculate the occurrence of specific words (prep. and conj.) , and classify data by levels. Obviously, beginner has many differences than others. "of" , "and" and "in" are Top 3 popular words.
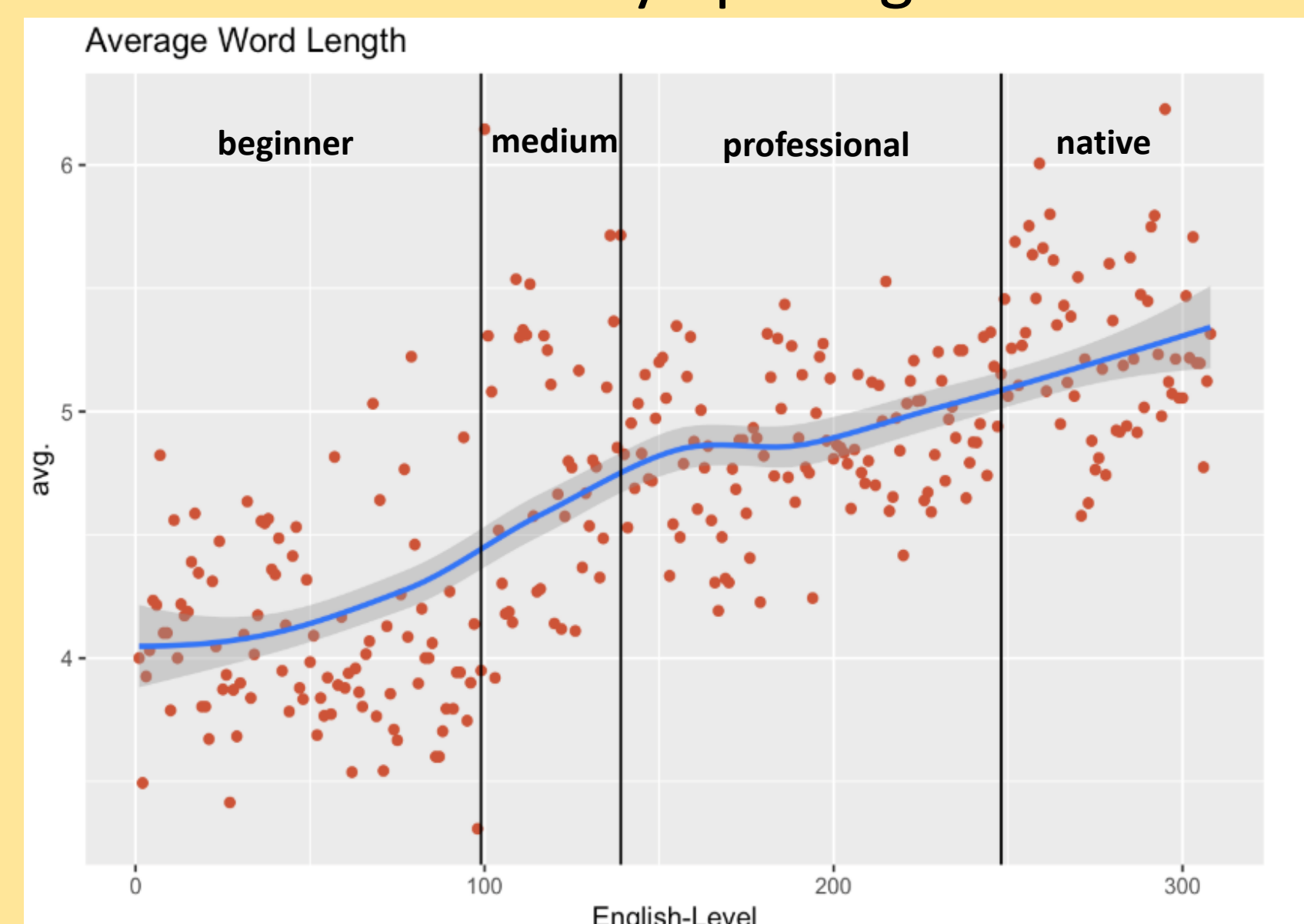

String Frequency by Language Levels

Lexical diversity

Calculate the lexical diversity in every English level. The lexical diversity means the ratio of difference words to the total number of the words I each article.
This chart shows that: The lower level a paragraph is, the higher in lexical diversity. We suppose that low-level-English users are used to use less words to write a article. Therefore, the rate end up in a greater number.

Average Word Length

Count the average word length of every English level. A slight tendency shows that: The higher level a paragraph is, the longer the avg. word length is. We assume that high-level-English user tend to pick longer words rather than easy-spelling short words.

VSM grader

To estimate the writer degree, we build a model with tf-idf vector space model.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad idf_i = \frac{|D|}{|\{j : t_i \in d_j\}|}$$

tf means the words repeatedly appear in a document and be counterbalance by total number of words. idf decrease the importance of a word if it appear in most of the document.
Our model accept an essay as input. It will return grade chart to evaluate the English degree of the essay.
We can estimate the popular words usage of write after calculating the document frequency. The work is shown as **Essay eye**

The result shows the beginner and medium level writers tend to focus on specific word usage while the professional or native writers search on diverse key word usage. The difference of native and professional write are little since similar skills.