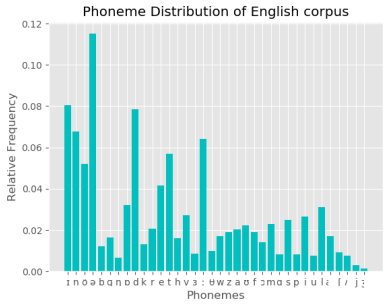
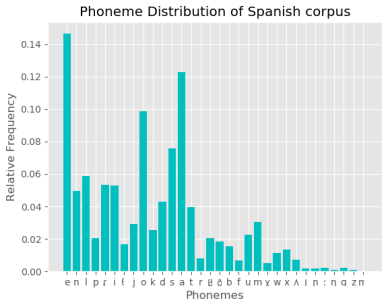
1 Kullback-Leibler Divergence

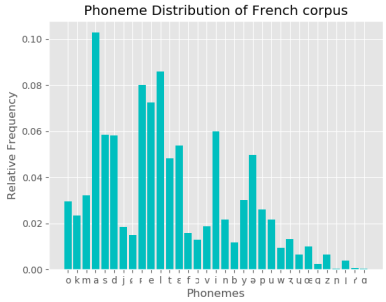
1.1 Phonemes and Language Similarity

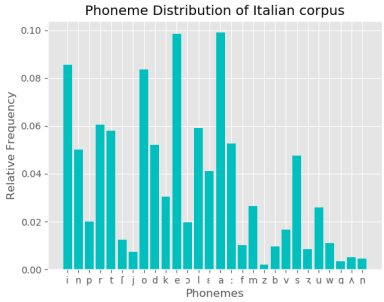
a) A unified phoneme set is constructed from all the different language phonemes. The probability distribution of phonemes is calculated for each corpus with Lidstone smoothing where, $\alpha = 1$. Then 3 test cases are run for different unseen phonemes in different corpus. The probability of unseen phoneme '$' is calculated in English and French corpus. The probability of another unseen phoneme '1' is calculated in Italian corpus. The test was successful, as in all cases it returns some very little probability value which is not zero.

b) From the bar chart below we can see that the distribution of phonemes in these 4 languages is not uniform at all. In each corpus frequency of some phonemes is much more than the others while there are also such phonemes which frequency is quite low than the remaining.









c) KL Divergence is computed for each corpus using the given formula which returns value in bits

d) From the table below we can analyze the closeness of the phonemes of different language pair from their respective probability distribution. Similarity between same corpus is 100% that is why in the table for the pair X*X the dissimilarity is 0. English phonemes are more closer to French than the others, Spanish are closer to Italian whereas, French is 1.6133101058269301 bits dissimilar to the Italian which is the closest from French to Italian. Italian is the closest to Spanish. However, we can not say vice versa. For example, we already analyzed that English phonemes are more closer to French than the others, but the opposite, French phonemes are more closer to English than the others does not hold. Because we already mentioned above that French phonemes have less dissimilarity with the Italian. So the value in the table English*French = 2.8261536675741468 bits and the value French*English = 2.4309552665208773 bits do not match. Hence, the asymmetric property of the KL-divergence is justified.

	English	Spanish	French	Italian
English	0.0	3.4527781919333425	2.8261536675741468	3.1871598716598757
Spanish	2.5457803433370376	0.0	1.3777115939406528	1.1271537967399012
French	2.4309552665208773	2.296411824471215	0.0	1.6133101058269301
Italian	1.8778719008184726	1.0740189099914261	1.1985568480541144	0.0

Table: KL Divergence for all language pair

2 Language Models Evaluation

2.1 LM Evaluation - Perplexity

a)

$$peplaxity(\mathcal{V}) = 2^{-log(\frac{1}{M})}$$

b)