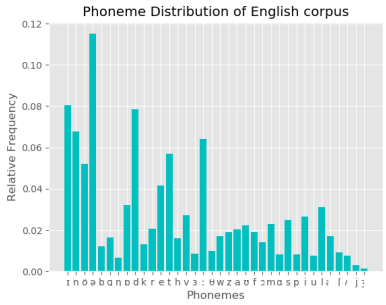
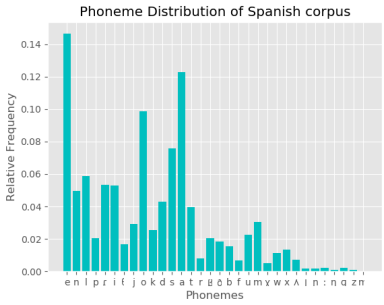
1 Kullback-Leibler Divergence

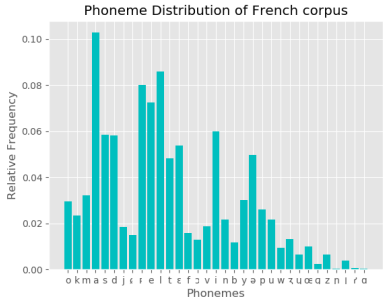
1.1 Phonemes and Language Similarity

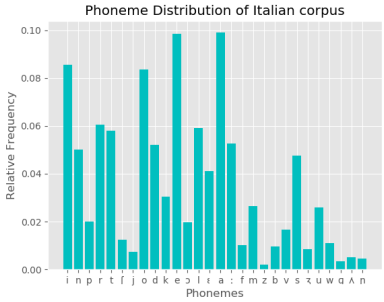
a) A unified phoneme set is constructed from all the different language phonemes. The probability distribution of phonemes is calculated for each corpus with Lidstone smoothing where, $\alpha = 1$. Then 3 test cases are run for different unseen phonemes in different corpus. The probability of unseen 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)

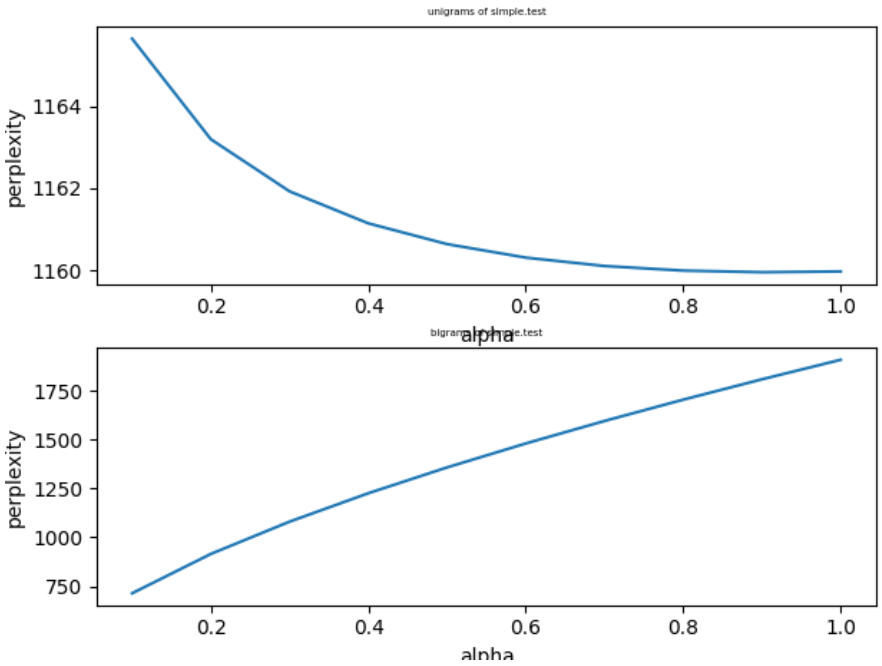
$$perplaxity(\mathcal{V}) = 2^{-log(\frac{1}{M})}$$

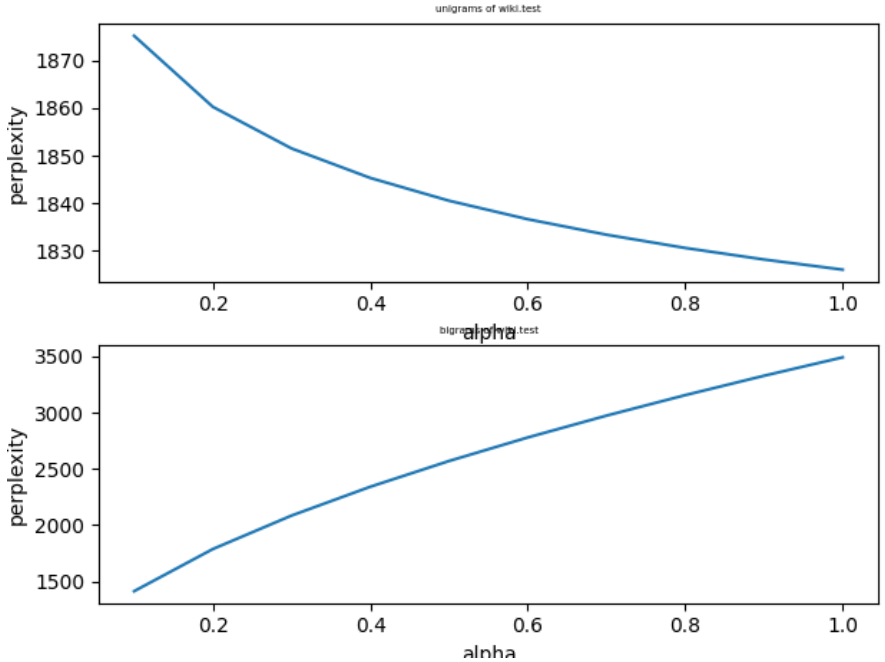
c) The table shows that the peplexity of the bigrams is in both cases lower that the perplexity of the unigrams. It was not possible to calculate the perplexity for alpha = 0 because, there are several words which are not in the learned vocabulary which results in a probability of 0 or a history of 0. Both cases lead to log(0) or divsion by zero which can not be calculated. For N untrained words with a theoretical very small probability we have N times $log(x \to 0)$ in our sum. This explaines the high values on these unknown test corpus.

simplet.test				wiki.test			
unigram		bigram		unigram		bigram	
smoothed	unsmoothed	smoothed	unsmoothed	smoothed	unsmoothed	smoothed	unsmoothed
1163	no result	915	no result	1860	no result	1788	no result

d) File simple.test has 1.96% unseen unigrams and 18.95% unseen bigrams. File wiki.test has 7.75% unseen unigrams and 31.42% unseen bigrams. The higher amount of unseen data in wiki.test explaines the higher perplexity values in the table of c), which is given by the fact that each time a new word is found the log of a realy small value gets added in the sum.

e) The plots show that a bigger alpha in the smoothing changes the perplexity of the bigrams. The unigrams dont change that much in comparission with the bigram values. The important thing is that while the uniram values drop with a higher alpha, the bigram values rise with bigger alpha. The scale of this changes is bound to the amount of unseen data in the test corpus.





f) In the book the difinition is 'a perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step.'
The definition on this exercise sheet is the sum of the logarithms of the individual likelihood for each word and then again the log gets reversed. So the definition would be something like the normalized avarage likelihood of all the words.

2.2 LM Evaluation – Grammatically Assessment

a) With score_sentence we computed the score for each Yoda-ish sentence and its English equivalent. The table of scores:

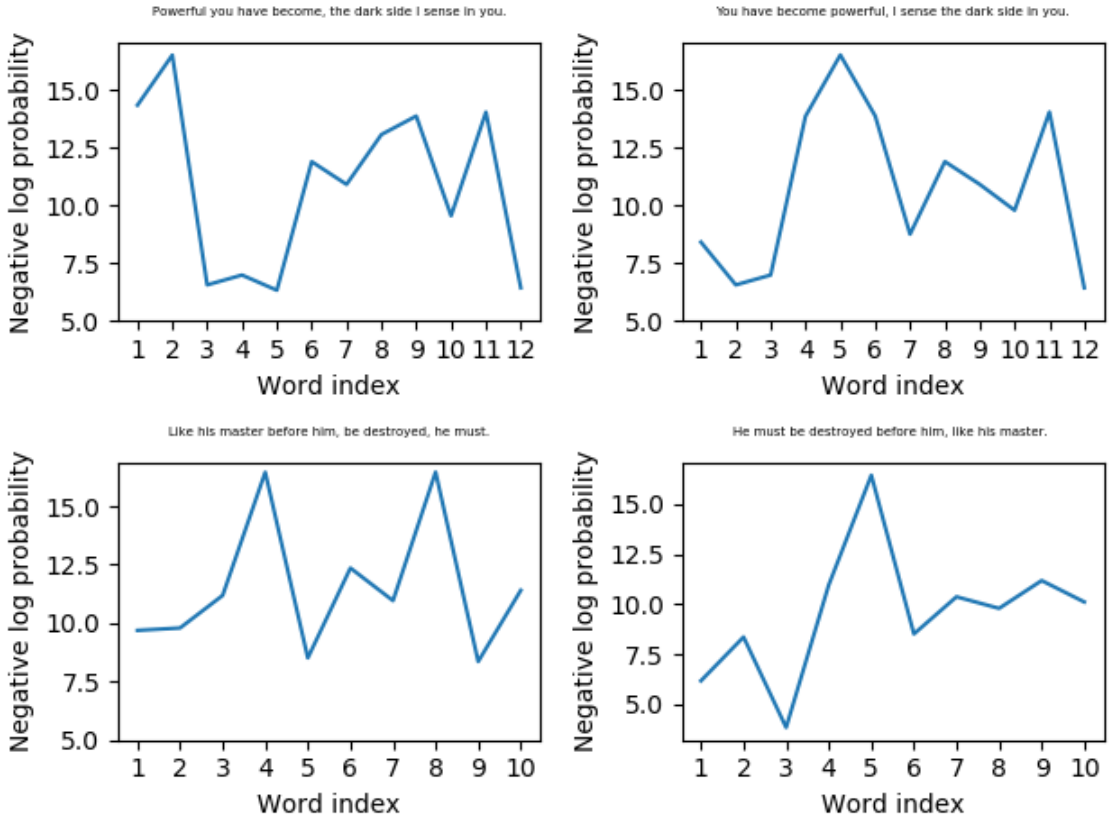
#	Yoda-ish sentence	Score_Yoda	Score_stand
1	Once you start down the dark path, forever will it dominate your destiny, consume you it will.	12.181	11.398
2	Patience you must have, my young padawan.	11.942	11.318
3	When nine hundred years old you reach, look as good you will not.	9.859	8.857
4	Truly wonderful, the mind of a child is.	10.115	8.853
5	Clear your mind must be, if you are to find the villains behind this plot.	10.487	9.898
6	Always two there are, no more, no less. A master and an apprentice.	9.219	8.887
7	In this war, a danger there is, of losing who we are.	8.713	7.927
8	So certain were you. Go back and closer you must look.	9.297	8.545
9	If no mistake have you made, yet losing you are ... a different game you should play.	9.926	8.542
10	Powerful you have become, the dark side I sense in you.	10.037	9.846
11	Through the Force, things you will see. Other places. The future…the past. Old friends long gone.	6.938	8.521
12	Size matters not. Look at me. Judge me by my size, do you?	9.446	8.847
13	You think Yoda stops teaching, just because his student does not want to hear? A teacher Yoda is. Yoda teaches like drunkards drink, like killers kill.	11.799	12.062
14	Happens to every guy sometimes this does.	11.379	10.361
15	The dark side clouds everything. Impossible to see the future is.	8.643	8.029
16	Soon will I rest, yes, forever sleep. Earned it I have. Twilight is upon me, soon night must fall.	11.336	9.841
17	Much to learn you still have ... my old padawan.	9.912	10.076
18	Like his master before him, be destroyed, he must.	10.471	8.705
19	Already know you that which you need.	8.772	8.381

Observations: mostly conventional English equivalents have lower value of the score, in other words, they are scored better than original Yoda-ish phrases.

However, there are several pairs where the scores have small difference and a few pairs where the score of Yoda's phrase is lower than the score of standard sentence.

b) Considering differences between scores of the sentences in each pair, we found the one with **the smallest difference: 'Powerful you have become, the dark side I sense in you.' / 'You have become powerful, I sense the dark side in you.'** ***dif_min = 0.191***

And the pair with **the largest difference: Like his master before him, be destroyed, he must.' / He must be destroyed before him, like his master.'** ***dif_max = 1.765***



Phrases in both pairs have graphs with a lot of rises and drops. In each pair, we can note those parts of the graphs that save its behavior; they correspond to those parts of sentences, which save the word order.

Larger score values of the Yoda-ish sentences is explained by those rises where the subject is replaced by other parts.

Typically, in graphs of standard phrases sharp rises denotes those words, which occur after punctuation (comma). Semantically they are beginning of the new phrase whereas they are scored in a sequence with the ending of the previous part.

It explains high score of standard phrase in the first pair, which became close to the score of Yoda's variant.

c) Bigram LM does not succeed well in the task of evaluating which of the phrases in the pair is more fluent. We can see that some of Yoda's phrases are scored to be better than conventional sentence structure. That might be caused by lower value of negative log probability of the words, which appear in the beginning of the sentence, because they often used in training corpus as a subject, whereas in Yoda's sentence they play role of an object.

d) Unigram LM cannot be applied in this task, as it does not represent any information regarding sentence structure and words sequences. It would give probabilities of single words in the sentence; therefore, scores for both phrases in each pair would be the same.

e) Russian is not as strict in sentence structure as English and usually words permutations does not make sentence less fluent. However in case the action is expressed by two verb (for example, 'You *want to teach* him' = 'Ты *хочешь обучать* его') moving of the main verb in the end of sentence would make it more unconventional, like from Old Russian epos ('You *to teach* him *want*' = 'Ты *обучать* его *хочешь*'). One more way: to change words in possessive phrases ('*The mind of a child* is truly wonderful.' → '*Of a child the mind* is truly wonderful.' In Russian: 'Воистину удивителен *разум ребёнка.*' → 'Воистину удивителен *ребёнка разум.*'). In order to strengthen the impression we could choose those translation options of words which look obsolete.