NLP Assignment

Clément Aguilar

Task 1

Tabulate, graph and explain your results.

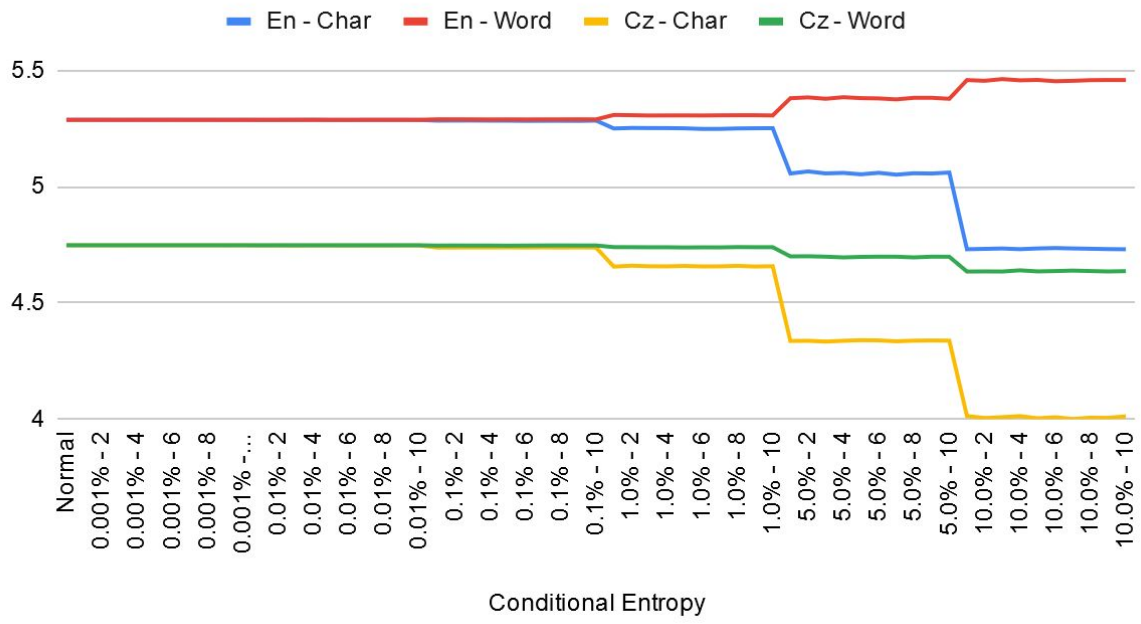
Conditional Entropy	En - Char	En - Word	Cz - Char	Cz - Word
Normal	5.287465249	5.287465249	4.747814095	4.747814095
0.001% - 1	5.287541823	5.287461449	4.747827072	4.747814095
0.001% - 2	5.287486804	5.287535901	4.74774922	4.747814095
0.001% - 3	5.287338593	5.287505126	4.747797526	4.747779593
0.001% - 4	5.287453256	5.287497884	4.747804982	4.747814095
0.001% - 5	5.287462434	5.287515028	4.747730529	4.747814095
0.001% - 6	5.287537098	5.287518196	4.747681896	4.747814095
0.001% - 7	5.287410078	5.2874765	4.747773873	4.747814095
0.001% - 8	5.287455562	5.287484016	4.747768176	4.747796757
0.001% - 9	5.2873165	5.28746282	4.747579611	4.747759649
0.001% - 10	5.287504922	5.28743967	4.747691043	4.747814095
0.01% - 1	5.286891169	5.287728769	4.746953717	4.747651614
0.01% - 2	5.287108947	5.287567208	4.746845397	4.747823689
0.01% - 3	5.28712299	5.287777844	4.746475562	4.747800993
0.01% - 4	5.287772629	5.287538978	4.747017853	4.747618073
0.01% - 5	5.286849514	5.287616097	4.746864957	4.747692484
0.01% - 6	5.286764936	5.287582195	4.746730852	4.747673837
0.01% - 7	5.287102998	5.287706221	4.74693249	4.747710749
0.01% - 8	5.286947904	5.287761681	4.746760089	4.747731581
0.01% - 9	5.287276305	5.287596549	4.746850146	4.747840404
0.01% - 10	5.286921912	5.287650149	4.746905153	4.747803168
0.01% - 1	5.284101993	5.28972482	4.737784489	4.746438751
0.1% - 2	5.284288765	5.289720212	4.738005046	4.746923835
0.1% - 3	5.284492974	5.289468884	4.738364828	4.746612456
0.1% - 4	5.283924806	5.28919498	4.73810906	4.746640836
0.1% - 5	5.283821076	5.289472261	4.738560835	4.746116154
0.1% - 6	5.283048979	5.289169464	4.737916369	4.746563957

0.1% - 7	5.283252916	5.289457453	4.738785984	4.746988798
0.1% - 8	5.283443022	5.289510681	4.737808509	4.747121487
0.1% - 9	5.283037846	5.289755328	4.738632448	4.746898719
0.1% - 10	5.284306673	5.289385226	4.737725571	4.746979006
1.0% - 1	5.250356674	5.308548215	4.655956487	4.739801214
1.0% - 2	5.252589901	5.307446003	4.66000028	4.739614924
1.0% - 3	5.252078174	5.306202062	4.657486205	4.739218517
1.0% - 4	5.251941876	5.306493855	4.657113636	4.739217326
1.0% - 5	5.251018641	5.306636217	4.658814666	4.738287126
1.0% - 6	5.248356713	5.306236643	4.656719047	4.738835747
1.0% - 7	5.248507158	5.30692821	4.657038834	4.738672782
1.0% - 8	5.250488671	5.307292578	4.65921888	4.740001225
1.0% - 9	5.251212312	5.307324919	4.656366558	4.739532234
1.0% - 10	5.251670179	5.306343368	4.657653061	4.739546016
5.0% - 1	5.056584769	5.380416606	4.336174145	4.69992948
5.0% - 2	5.065531665	5.384051075	4.336878045	4.700489564
5.0% - 3	5.056952085	5.378121823	4.333263309	4.698700097
5.0% - 4	5.059185292	5.384811706	4.336891517	4.695558441
5.0% - 5	5.053061932	5.380828216	4.339394824	4.697752968
5.0% - 6	5.059595055	5.379637452	4.338590373	4.698733341
5.0% - 7	5.051591523	5.375477025	4.334647945	4.698584189
5.0% - 8	5.057611687	5.38228803	4.33736614	4.695619798
5.0% - 9	5.056505918	5.382330599	4.338368826	4.698726445
5.0% - 10	5.060549115	5.378002003	4.337442179	4.698371528
10.0% - 1	4.73071944	5.458558413	4.013213829	4.63471236
10.0% - 2	4.732284313	5.455429402	4.004544398	4.635229002
10.0% - 3	4.733930907	5.462803346	4.008076385	4.634821187
10.0% - 4	4.730978873	5.457517579	4.012687688	4.639975844
10.0% - 5	4.734178795	5.459124831	4.003223166	4.635373054
10.0% - 6	4.735830581	5.453632544	4.00797352	4.636772648
10.0% - 7	4.734168602	5.455363231	4.000354377	4.638794431
10.0% - 8	4.732789252	5.458238299	4.006187526	4.636854596
10.0% - 9	4.731372615	5.458835246	4.00513656	4.635097532
10.0% - 10	4.730527641	5.458638635	4.011798122	4.636420844

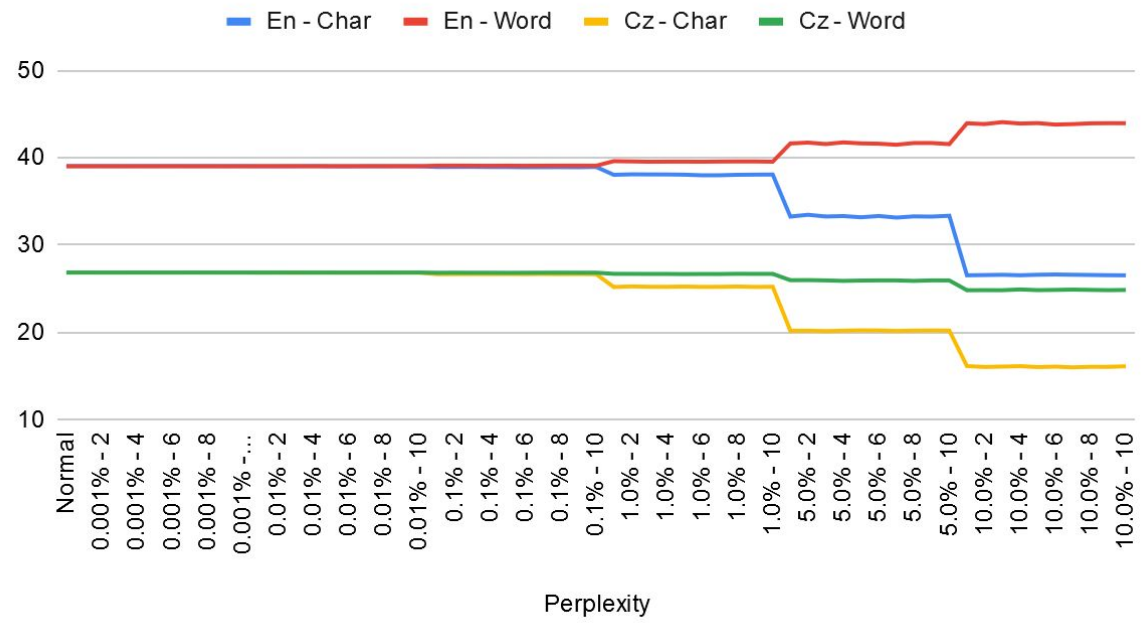
Perplexity	En - Char	En - Word	Cz - Char	Cz - Word
Normal	39.05580925	39.05580925	26.86794536	26.86794536
0.001% - 1	39.05788227	39.05570637	26.86818704	26.86794536
0.001% - 2	39.05639278	39.05772195	26.86673719	26.86794536
0.001% - 3	39.05238066	39.0568888	26.8676368	26.86730282
0.001% - 4	39.05548458	39.05669273	26.86777565	26.86794536
0.001% - 5	39.05573304	39.05715687	26.86638912	26.86794536
0.001% - 6	39.05775435	39.05724264	26.86548347	26.86794536
0.001% - 7	39.05431572	39.05611384	26.86719629	26.86794536
0.001% - 8	39.05554702	39.05631731	26.8670902	26.86762247
0.001% - 9	39.05178262	39.0557435	26.86357882	26.8669314
0.001% - 10	39.05688328	39.05511679	26.86565381	26.86794536
0.01% - 1	39.04027118	39.06294377	26.85192694	26.86491958
0.01% - 2	39.04616484	39.05856953	26.84991093	26.86812403
0.01% - 3	39.04654492	39.06427255	26.84302884	26.86770136
0.01% - 4	39.06413134	39.05780526	26.85312069	26.86429501
0.01% - 5	39.03914398	39.05989313	26.85027495	26.86568064
0.01% - 6	39.03685539	39.05897527	26.84777921	26.8653334
0.01% - 7	39.04600385	39.06233324	26.85153186	26.86602077
0.01% - 8	39.0418065	39.06383491	26.84832332	26.86640872
0.01% - 9	39.05069462	39.0593639	26.84999931	26.86843533
0.01% - 10	39.04110312	39.06081508	26.85102306	26.86774186
0.01% - 1	38.96486718	39.11702698	26.68180736	26.84234392
0.1% - 2	38.96991192	39.11690204	26.68588675	26.85137076
0.1% - 3	38.97542838	39.11008819	26.69254255	26.84557603
0.1% - 4	38.96008193	39.10266361	26.68781078	26.84610413
0.1% - 5	38.95728079	39.11017974	26.69616928	26.83634245
0.1% - 6	38.93643734	39.10197205	26.6842465	26.84467357
0.1% - 7	38.9419417	39.10977831	26.70033584	26.85257988
0.1% - 8	38.94707348	39.11122127	26.68225159	26.85504971
0.1% - 9	38.93613686	39.11785417	26.69749447	26.85090331
0.1% - 10	38.97039563	39.10782036	26.68071772	26.85239763
1.0% - 1	38.06403699	39.63074587	25.21056397	26.71913159
1.0% - 2	38.12300403	39.60047974	25.28132689	26.71568166
1.0% - 3	38.10948412	39.56634958	25.2373094	26.70834206
1.0% - 4	38.10588392	39.57435289	25.23079284	26.70832001
1.0% - 5	38.08150633	39.57825819	25.26055911	26.69110495

1.0% - 6	38.01130663	39.56729797	25.22389294	26.70125685
1.0% - 7	38.01527068	39.58626941	25.22948468	26.69824088
1.0% - 8	38.06751977	39.5962686	25.26763759	26.72283611
1.0% - 9	38.08661883	39.59715626	25.21773082	26.71415047
1.0% - 10	38.09870827	39.57022512	25.24022841	26.71440567
5.0% - 1	33.28002863	41.65496631	20.19847049	25.99080621
5.0% - 2	33.48705651	41.7600367	20.20832785	26.00089833
5.0% - 3	33.28850292	41.58876166	20.15775837	25.96866774
5.0% - 4	33.34007148	41.78205962	20.20851657	25.91217911
5.0% - 5	33.1988631	41.66685242	20.24361201	25.95162491
5.0% - 6	33.34954227	41.63247585	20.23232727	25.96926614
5.0% - 7	33.16504374	41.51258955	20.1771142	25.96658148
5.0% - 8	33.30372596	41.70903504	20.21516593	25.91328118
5.0% - 9	33.27820974	41.71026576	20.22922053	25.96914202
5.0% - 10	33.37160374	41.58530772	20.21623142	25.96275414
10.0% - 1	26.55146279	43.97337662	16.14721922	24.84205057
10.0% - 2	26.58027844	43.87810764	16.05047845	24.85094832
10.0% - 3	26.6106327	44.10295281	16.08982115	24.84392456
10.0% - 4	26.55623784	43.94166341	16.14133152	24.93284907
10.0% - 5	26.6152054	43.99064445	16.03578601	24.85342981
10.0% - 6	26.64569541	43.82349204	16.08867399	24.87755243
10.0% - 7	26.61501735	43.87609517	16.00393065	24.9124401
10.0% - 8	26.58958309	43.96362062	16.06876922	24.87896555
10.0% - 9	26.56348659	43.98181531	16.0570678	24.84868382
10.0% - 10	26.54793314	43.97582187	16.13138184	24.87148673

Conditional Entropy



Perplexity



Aside from the curve representing the experiment where I replaced the English words randomly by other words, the results seem to indicate the same thing.

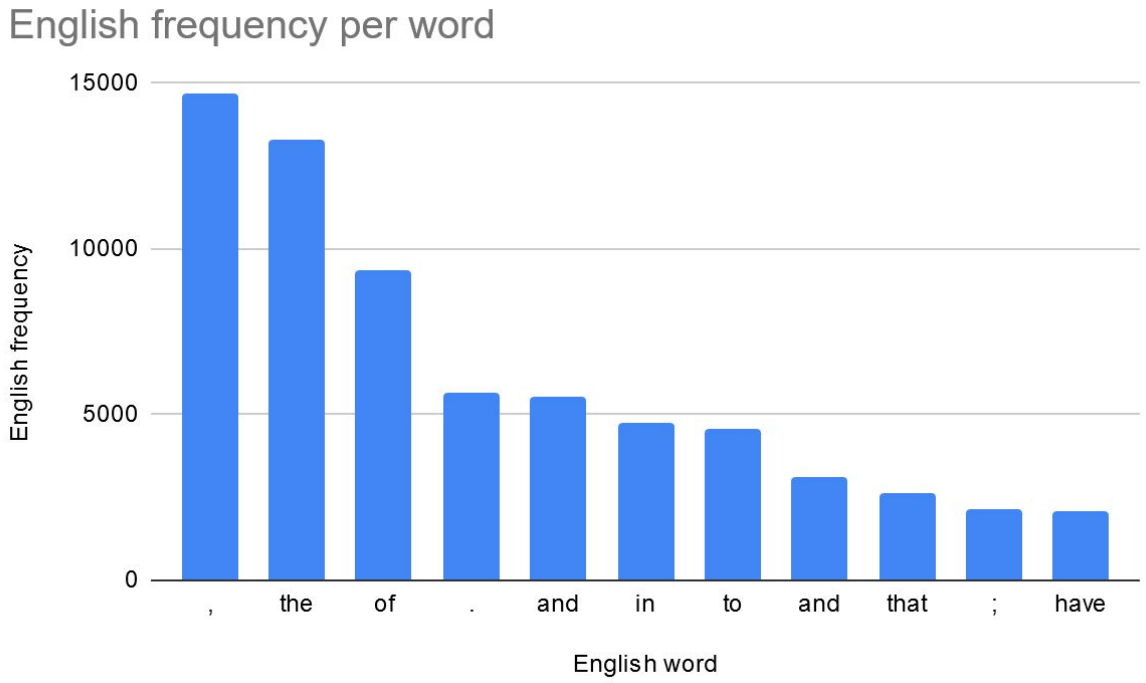
There are several observations we can make:

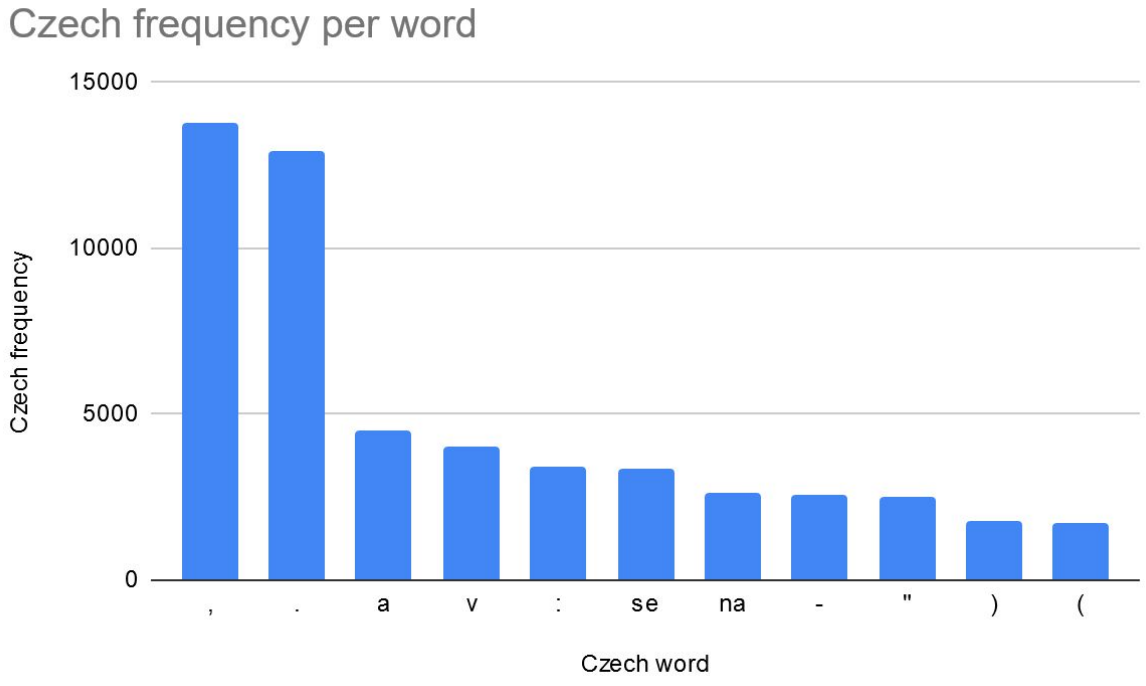
- Conditional entropy and perplexity are completely proportional (no surprise)
- The effect of changing characters is much stronger than the effect of changing words, as changing just characters will create new words, which will have drastic effects on the characteristics of the text. This can be clearly seen from the graph, as the two curves for the character-changing mess-ups move a lot more than those for the word-changing mess-ups.
- Surprisingly, one of the curves, the one for the word-changing mess-ups in English, goes up instead of going down like the others. I will discuss this in the next part.
- Mess-ups from 0.001% to 0.1% have little to no impact on conditional entropy, meaning we can afford messing up a text a bit without completely changing its characteristics.
- However, mess-ups starting from 1.0% have quite a substantial impact on the results and this only goes increasing.

Also try to explain the differences between the two languages. To substantiate your explanations, you might want to tabulate also the basic characteristics of the two texts, such as the word count, number of characters (total, per word), the frequency of the most frequent words, the number of words with frequency 1, etc.

Top 11 frequent words per language:

English word	English frequency	Czech word	Czech frequency
,	14721	,	13788
the	13299	.	12931
of	9368	a	4486
.	5645	v	4043
and	5537	:	3434
in	4761	se	3378
to	4548	na	2646
and	3132	-	2549
that	2637	"	2506
;	2151	)	1761
have	2084	(	1748


English frequency per word
15000
10000
5000
0
English frequency
, the of . and in to and that ; have
English word


Czech frequency per word
15000
10000
5000
0
Czech frequency
, . a v : se na - ") (
Czech word

Number of English words: 221098 / distinct words: 9607

Number of Czech words: 222412 / distinct words: 42826

Number of English characters: 972917

Number of Czech characters: 1030631

Number of words with frequency 1 in the English text: 3811

Number of words with frequency 1 in the Czech text: 26315

Here we can see interesting differences between the two languages that have to do with the results we are observing. Here are my thoughts on this:

- The sudden rise in conditional entropy of word-changing mess-ups in English might be due to a computer problem. After seeing the way the conditional entropy was changing, I had a long look at the program, looking for a mistake, but it was almost the same as the one for the character-changing mess-ups in English which did not show any anomaly.
- Character-changing mess-ups have similar effects on both languages, which makes sense. It basically does the same thing in both languages: creating new words. However, word-changing mess-ups have different effects on the two languages, which is mainly due to the completely different structure of the two languages.
- The frequency of the punctuation in the two languages is not the same. This is so true that Czech almost looks like an exception to Zipf's law, as the two most frequent words are so close to one another.
- There are many more distinct words in Czech compared to English, which is largely a consequence of declension.
- Czech declension also plays a role in the huge difference in the number of words with frequency 1 between the two languages (almost 7 times more for Czech!).

Attach your source code commented in such a way that it is sufficient to read the comments to understand what you have done and how you have done it.

The source code has been commented in order to be understandable. The language that was chosen for the assignment is Scala.

Now assume two languages, L1 and L2 do not share any vocabulary items, and that the conditional entropy as described above of a text T1 in language L1 is E and that the conditional entropy of a text T2 in language L2 is also E. Now make a new text by appending T2 to the end of T1. Will the conditional entropy of this new text be greater than, equal to, or less than E? Explain (This is a paper-and-pencil exercise of course!)

Let's call the new text T3. In my opinion, its conditional entropy will be a tiny bit smaller than E, almost equal to it. This is because of the bigrams at the end and beginning of the texts.

For the text “a b b”, there are actually four bigrams, not two:
(_, a), (a, b), (b, b) and (b, _) where _ is the empty word.

When we append the two texts, the end bigram (b, _) will be replaced by a new bigram, the second element of which being the first word of the second text.
This has the result of slightly decreasing the overall conditional entropy of the new text since it removes a word from the total word count.

If it was not for that little detail, the conditional entropy of T3 would be equal to E. After all, the words from T1 do not appear in T2 and vice-versa, so the probabilities p(i|j) would not change. The only probabilities that would change are p(i,j) which would be divided by two since the text is now twice as long as before.

The conditional entropy would then have been ½ E + ½ E = E.

Task 2

Now compute the four smoothing parameters (i.e. "coefficients", "weights", "lambdas", "interpolation parameters" or whatever, for the trigram, bigram, unigram and uniform distributions) from the heldout data using the EM algorithm. (Then do the same using the training data again: what smoothing coefficients have you got? After answering this question, throw them away!)

Note: for the following results and for the rest of the assignment, I systematically used the value 0.00000001 for epsilon.

Parameters for the held-out data:
L0 = 0.044347262170143044
L1 = 0.5658670070285023
L2 = 0.36875597928064113
L3 = 0.021029751520713503

Parameters for the training data:
L0 = 3.1195632553544437E-180
L1 = 2.912401740790397E-17
L2 = 0.031071694522390347
L3 = 0.9689283054776096

The results on the training data make sense with the indication provided on page 87 of the slides: l3 is indeed converging towards 1.

And finally, compute the cross-entropy of the test data using your newly built, smoothed language model.

Cross-Entropy	EN	CZ
Normal	3.931844133	5.306206052

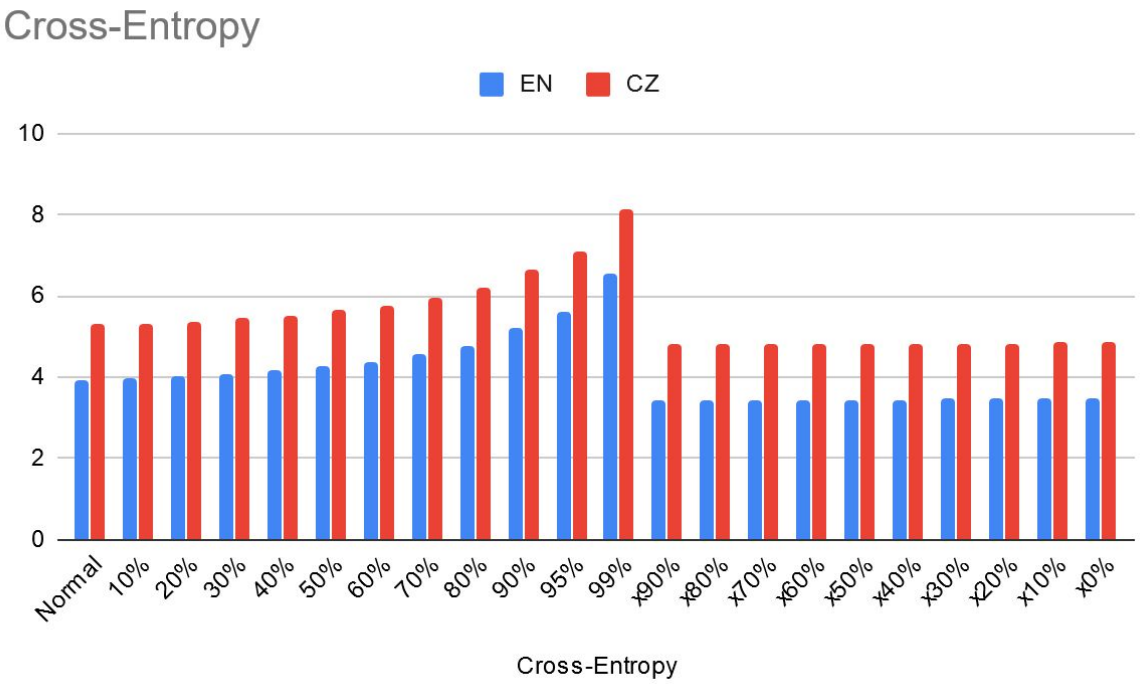
Now tweak the smoothing parameters in the following way: add 10%, 20%, 30%, ..., 90%, 95% and 99% of the difference between the trigram smoothing parameter and 1.0 to its value, discounting at the same the remaining three parameters proportionally (remember, they have to sum up to 1.0!!). Compute the cross-entropy on the test data for all these 22 cases (original + 11 trigram parameter increase + 10 trigram smoothing parameter decrease).

Cross-Entropy	EN	CZ
10%	3.949295371	5.318995322
20%	3.998582536	5.372195874
30%	4.064615783	5.443345128
40%	4.146892212	5.531597566
50%	4.248600173	5.640331805
60%	4.376670636	5.776972541
70%	4.544961216	5.956385563
80%	4.785116361	6.21258945
90%	5.198195808	6.654658033
95%	5.610871417	7.098650923
99%	6.561971569	8.13089642

Then set the trigram smoothing parameter to 90%, 80%, 70%, ... 10%, 0% of its value, boosting proportionally the other three parameters, again to sum up to one. Compute the cross-entropy on the test data for all these 22 cases (original + 11 trigram parameter increase + 10 trigram smoothing parameter decrease).

Cross-Entropy	EN	CZ
x90%	3.445167821	4.821321606
x80%	3.446076976	4.822737698
x70%	3.447090524	4.824321922
x60%	3.448229229	4.826114151
x50%	3.449522103	4.828172292
x40%	3.45101205	4.83058607
x30%	3.452767901	4.833508131
x20%	3.454915175	4.837239637
x10%	3.457744209	4.842567415
x0%	3.46279446	4.855693413

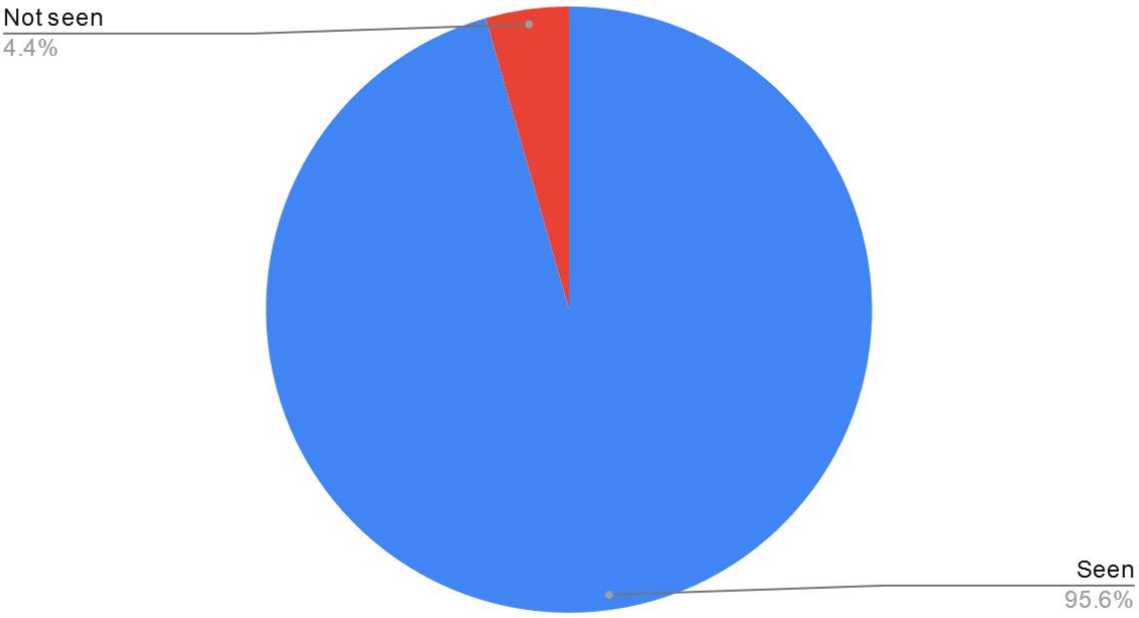
Tabulate, graph and explain what you have got. Also, try to explain the differences between the two languages based on similar statistics as in the Task No. 2, plus the "coverage" graph (defined as the percentage of words in the test data which have been seen in the training data).



This graph shows the effects of tweaking L3. Here are a few observations we can make:

- Increasing the weight of L3 seems to have immediate, very bad consequences on the cross-entropy. This is due to a number of reasons, two of which being:
    - Not all trigrams from the training data appear in the test data. If we only listen to the trigrams and ignore the bigrams and unigrams, a big chunk of the test data will be very unknown to our model.
    - When increasing L3, we decrease the role of the other parameters, so they cannot step in and try to correct the mistakes that are being made.
- Decreasing the weight of L3 also has bad consequences, but the loss in accuracy is very small when compared to the increase. This is due to the fact that as we decrease the role of the trigrams in our model, the unigrams, and especially the bigrams, will step up. They will gradually replace the role of the trigrams, leading to a small loss in accuracy.
- The difference between the languages has again a lot to do with Czech declension and conjugation, many times more complex than English. This can be seen from the coverage graphs (you can see them on the next page). A big difference in the number of unseen words can clearly be noticed and this will have quite the impact on cross-entropy.
- It could be interesting to lemmatize the two documents and see if those differences persist, as this would eliminate, at least in part, the differences in grammar complexity between the two languages.

English Coverage Graph



Czech Coverage Graph