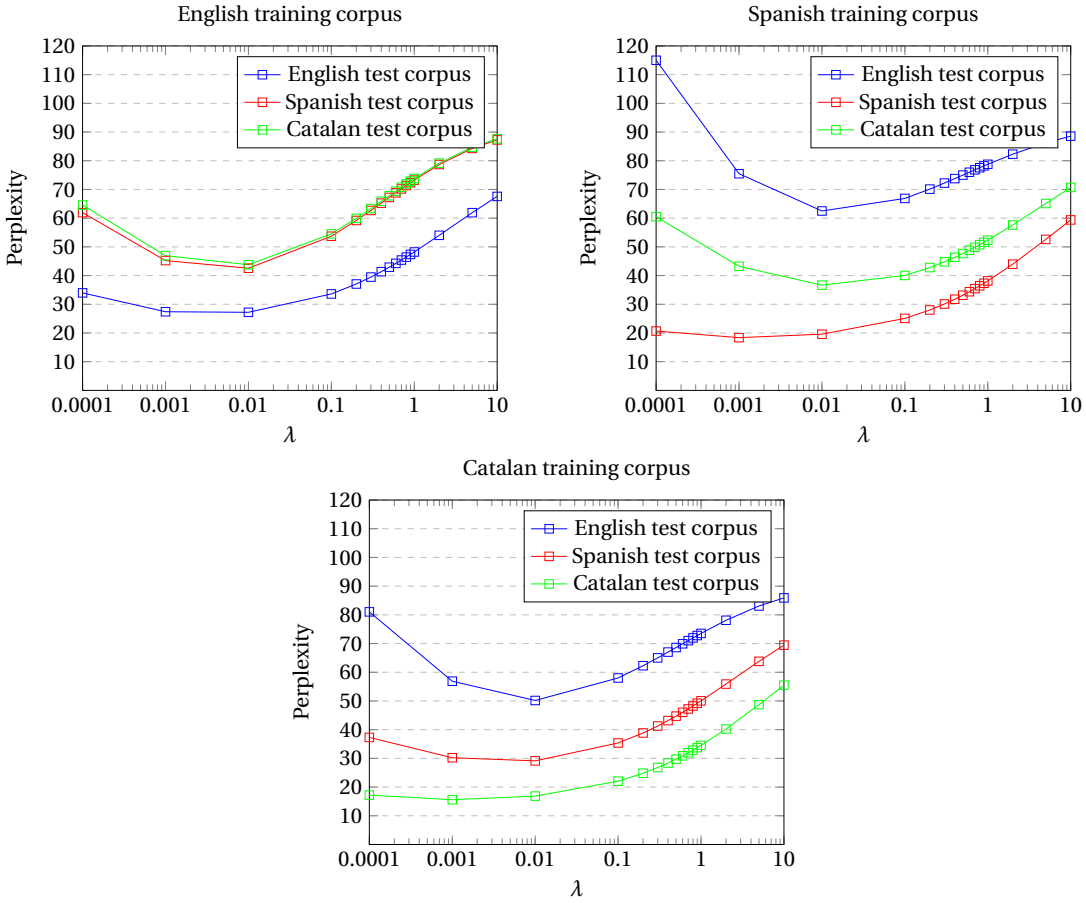
Natural Language Processing - Assignment 1

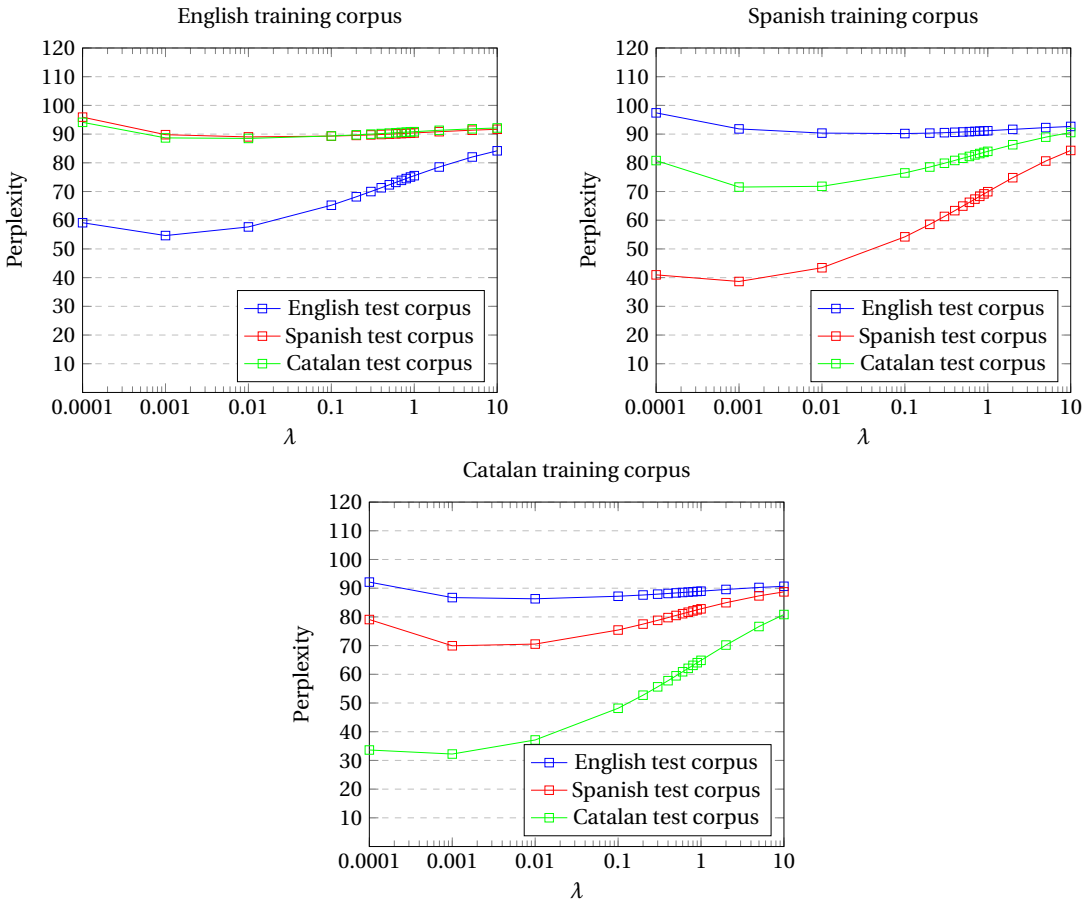
Oran Avraham (ID. 203539598) Amit Shaked (ID. 203119417)

1 Perplexity vs. λ-values in Lidstone's smoothing

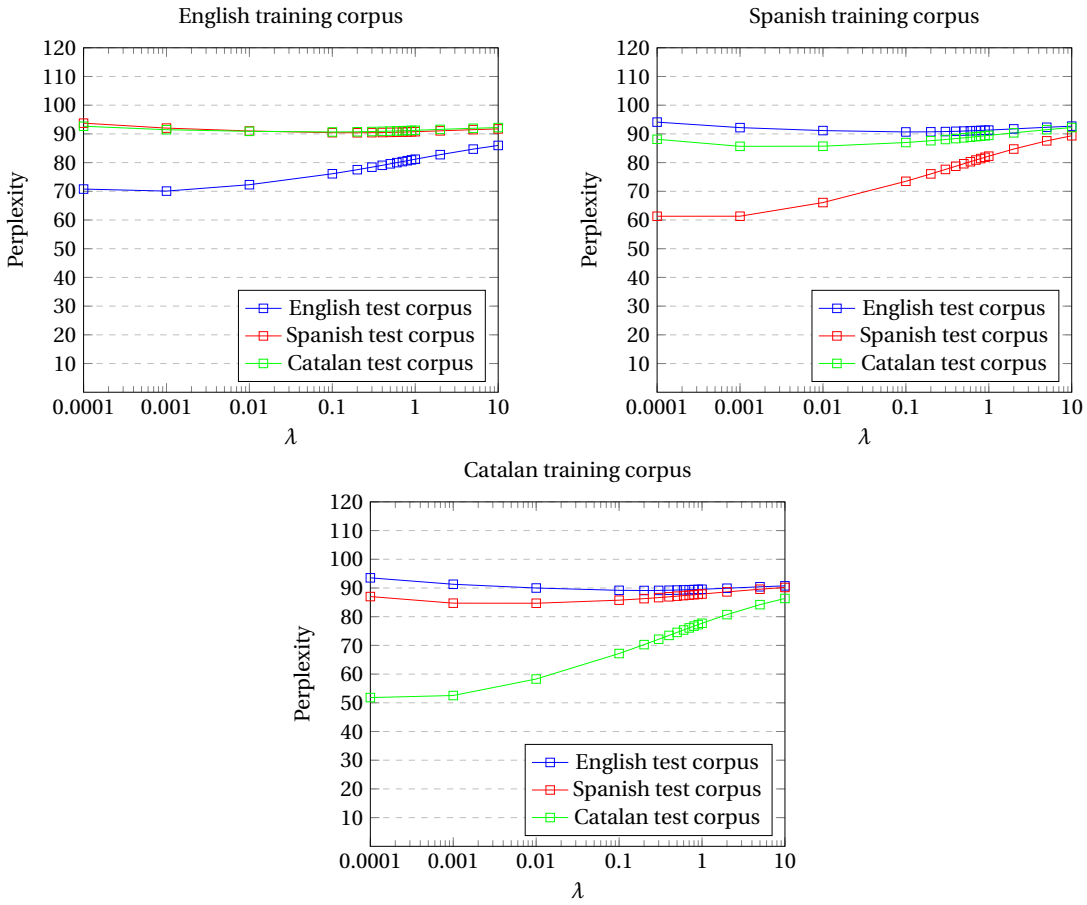
N=2



N=3


English training corpus
Spanish training corpus
Catalan training corpus
Perplexity
λ
English test corpus
Spanish test corpus
Catalan test corpus
120
110
100
90
80
70
60
50
40
30
20
10
0.0001
0.001
0.01
0.1
1
10

N=4

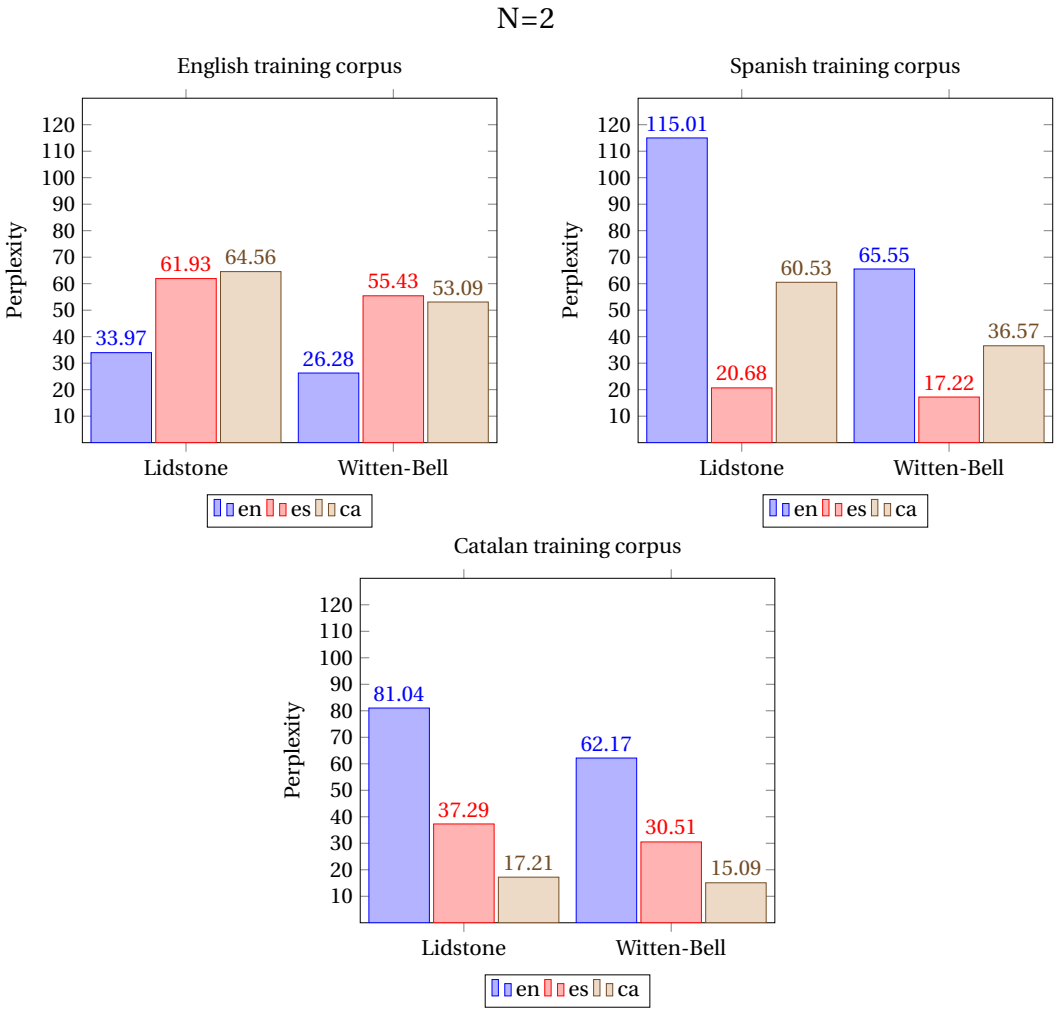


CONCLUSIONS

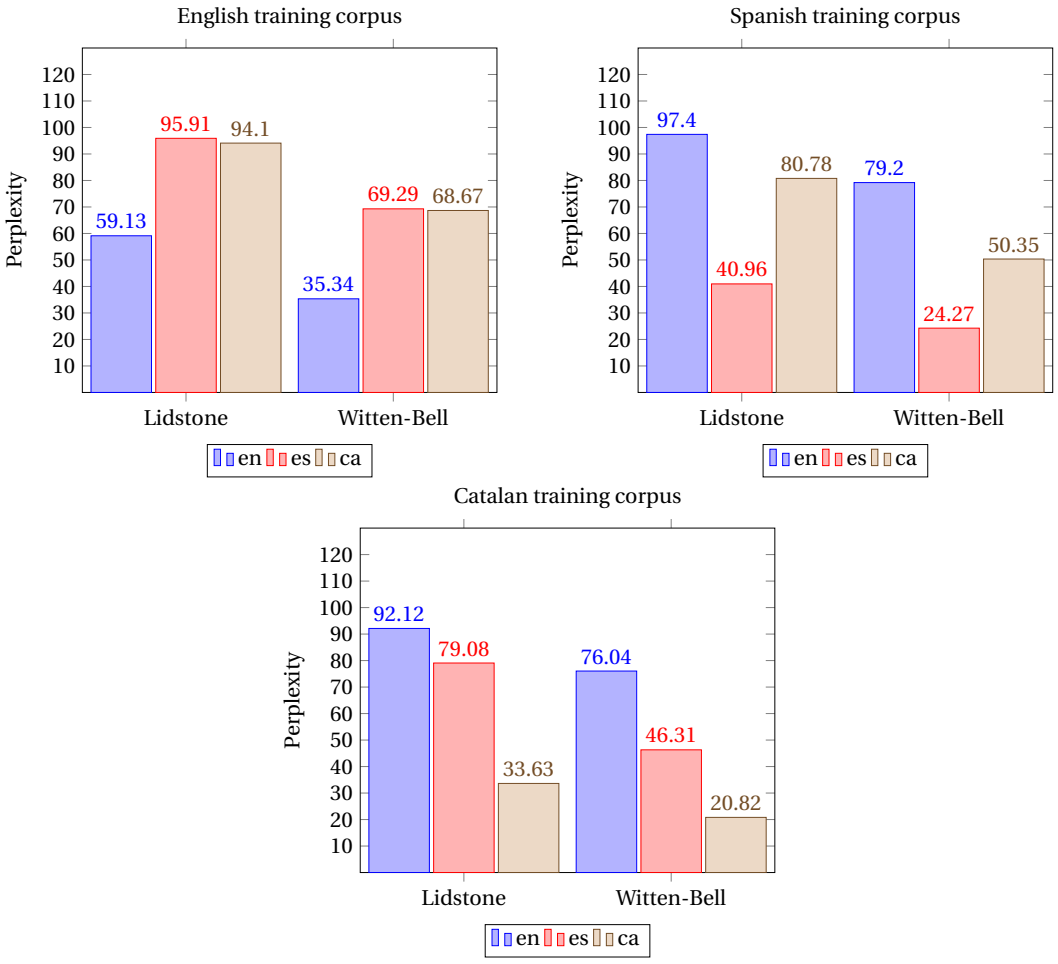
- Larger n-gram means larger perplexity values.
- Larger λ-values might mean smaller perplexity values for very small λ-values, but for $\lambda \geq 0.01$ the perplexity increases as λ increases (the graph is convex).
- As λ increases, the difference between condition positive (true language) perplexity and condition negative (false language) perplexity decreases (this is more noticeable for two-grams and three-grams). This might be due to the fact that larger λ-values means that the distribution is closer to uniform distribution.
- In our case, $\lambda = 10^{-3}$ seems to be the value for which the perplexity is minimal for the true test language. However, the difference in the perplexity of the true language between $\lambda = 10^{-4}$ and $\lambda = 10^{-3}$ is negligible, while for false languages there is a large difference. Hence, since our purpose is language detection, we chose $\lambda = 10^{-4}$.

2 WITTEN-BELL COMPARED TO LIDSTONE'S SMOOTHING

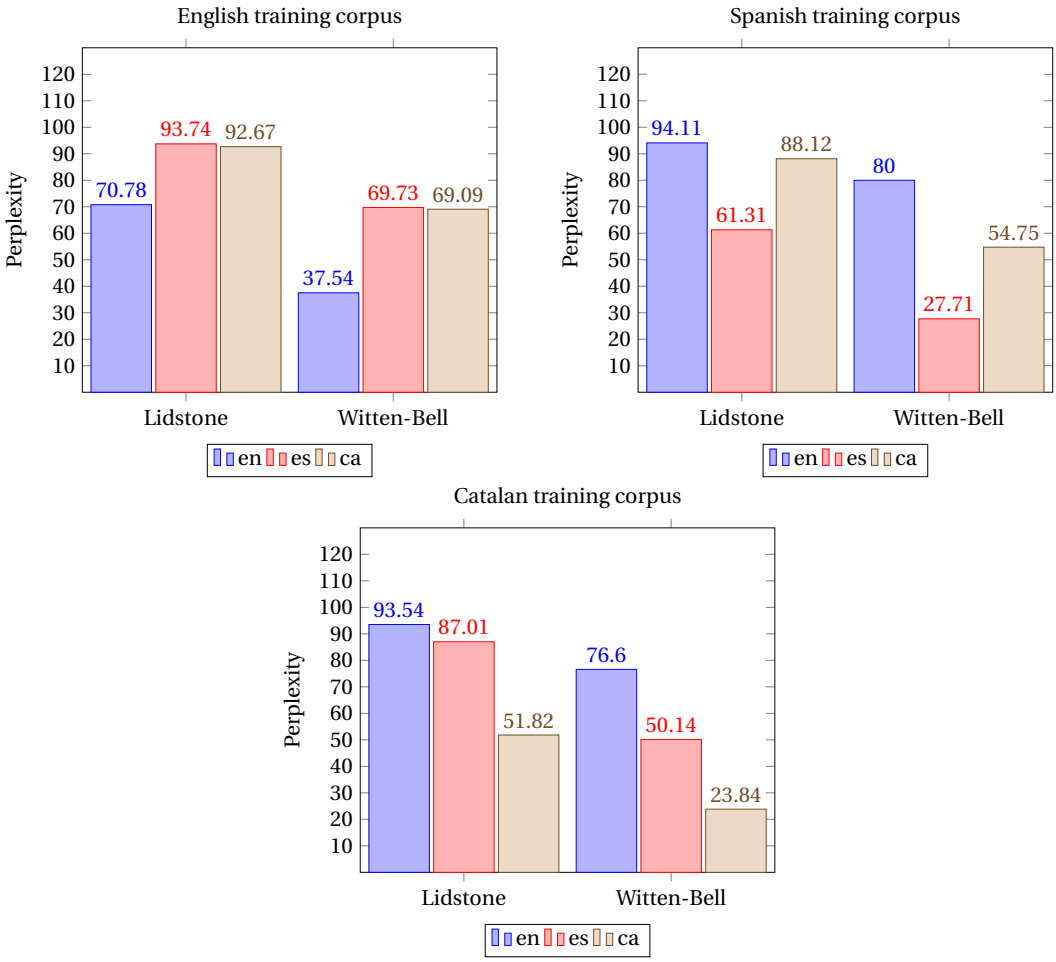
We chose $\lambda = 10^{-4}$ for Lidstone's smoothing as explained in previous section.



N=3


English training corpus
Perplexity
120
110
100
90
80
70
60
50
40
30
20
10
59.13
95.91
94.1
35.34
69.29
68.67
Lidstone
Witten-Bell
en es ca
Spanish training corpus
Perplexity
97.4
40.96
80.78
79.2
24.27
50.35
Lidstone
Witten-Bell
en es ca
Catalan training corpus
Perplexity
92.12
79.08
33.63
76.04
46.31
20.82
Lidstone
Witten-Bell
en es ca

N=4



CONCLUSIONS

- Witten-Bell smoothing generally produces lower perplexity values than Lidstone's.
- As n increases, the increase in perplexity values in Lidstone's smoothing is much larger than in Witten-Bell smoothing.
- For four-grams, Lidstone's smoothing showed large difference in perplexity values between Spanish and Catalan; Witten-Bell, however, showed smaller difference. This may be due to the fact that we incorporated interpolation in Witten-Bell, thus smaller n-grams probabilities are also considered.

3 LANGUAGE DETECTION

In the previous section we've seen that Witten-Bell smoothing produces lower perplexity values than Lidstone's, thus it models the language better. We would like to find the language model which suits our needs the best.
For two-grams, Spanish and Catalan seem to be too close to one another (when a Spanish corpus was tested against a Catalan model, there was a 15.42 difference in perplexity).
The results for three-grams and four-grams seemed to be somewhat similar; however, three-grams had lower perplexity values for the real language. In addition, as we've seen, four-grams is prone to overfitting (it models the training corpus instead of the language). Therefore, we chose $n = 3$ with Witten-Bell smoothing.
We've detected correctly 95% of the sentences: 94% of the English sentences, 96.5% of the Spanish sentences and 95% of the Catalan sentences.
Whenever we failed to identify a Spanish sentence, we identified it as Catalan. The opposite wasn't true (we did identify Catalan sentences as English).
Besides the language model, another option to implement language detection is to use a lexicon - for each language, store a lexicon of known words in this language; process the test corpus and find which lexicon covers the maximum number of words in the test corpus.