Yi En Gan (Andrew)
ECE 20875 – HW8 Problem 3

The ten most common n-grams for mystery.txt were as follows.
[(' de', 123), ('de ', 102), ('el ', 66), (' co', 56), ('os ', 56), ('as ', 53), ('do ', 49), (' la', 49), (' el', 45), ('ia ', 44)]

The ten most common n-grams for different language files.
English.txt
[('the', 149), (' th', 142), (' an', 129), ('he ', 121), ('nd ', 113), ('and', 111), ('ion', 102), (' of', 93), ('of ', 89), ('tio', 88)]

French.txt
[(' de', 204), ('es ', 183), ('de ', 136), (' et', 118), ('ion', 108), ('te ', 106), ('nt ', 100), ('e d', 98), ('et ', 93), ('tio', 92)]

German.txt
[('en ', 251), ('er ', 187), ('der', 152), (' un', 124), ('und', 122), ('nd ', 117), ('ein', 110), ('ung', 102), ('cht', 98), (' de', 94)]

Italian.txt
[(' di', 177), ('to ', 124), (' de', 99), ('la ', 98), (' in', 97), ('ion', 95), (' e ', 93), ('di ', 89), ('e d', 89), ('re ', 81)]

Portuguese.txt
[('os ', 138), ('de ', 134), (' de', 129), (' a ', 111), (' e ', 98), ('em ', 95), ('o d', 89), ('to ', 86), ('ao ', 78), (' di', 77)]

Spanish.txt
[(' de', 249), ('os ', 137), ('de ', 135), ('ion', 115), (' la', 104), ('cio', 101), ('la ', 97), (' y ', 92), (' a ', 84), ('_', 77)]

Mystery.txt might be French, Portuguese or Spanish.

By comparing the histograms of the language files with the mystery text, English, German, French and Italian can be eliminated from the pool of possible guesses. This is because the histogram for English.txt has a sharp spike at the 85, whereas the German.txt histogram has a spike at 22. The histogram for French.txt showed a high frequency at 83-84, which is not seen in Mystery.txt histogram. Italian has a very tall bar at 2-5.

The remaining possible guesses are Portuguese and Spanish. Although the histogram patterns for Portuguese and Spanish are like that of Mystery.txt, it is noted that the frequency count for Portuguese n-grams are a lot higher at 0 and 25, whereas the frequency for Spanish n-grams are very similar to the height of Mystery.txt n-grams.

It is believed that the language contained in the Mystery.txt is Spanish.