Comparing Is-English Programs for Use in Cracking Shift and Affine Ciphers

Pranav Boreddy

June 25, 2025

1 Introduction

In this project, I tested how well two common methods could decrypt classic substitution ciphers: the Caesar Cipher and the Affine Cipher. These ciphers are some of the oldest and most basic ways to hide messages by replacing letters using simple math rules. While they're not used anymore for real security, they're useful for learning about how encryption works.

The challenge was that the messages weren't just plain English but rather, they were made up of half English words and half gibberish. But the gibberish wasn't totally random. It was made to have the same letter frequencies as real English. That means methods that rely only on letter frequency might not work as well as expected.

Something to note is that a ciphertext produced by a Shift or Affine cipher would not naturally contain half English and half gibberish. This setup was intentionally designed as a pedagogical experiment, rather than a reflection of real-world ciphertexts. By mixing genuine English words with frequency-balanced gibberish, I created a controlled scenario that highlights the strengths and weaknesses of each method. While such a mixture may not occur naturally, it models a type of adversarial noise where statistical signals are misleading. In this environment, the Common Words Approach prevails because it directly identifies real content, whereas the Dot Product Approach struggles because it cannot distinguish between genuine words and deceptive text with English-like frequencies.

For both decryption methods, the text was preprocessed to have all spaces, punctuation, and capitalization removed. Both methods were fed with what is in a sense one large string.

I tested two decryption approaches:

1. **Common Words Approach:** This method checks how many real English words show up after trying each possible decryption. By iterating from i:i+len(word), I was able to iterate through each and every letter testing for that word for all words in the 10,000 most common english words dataset.

2. **Dot Product Approach:** This one used dot products to compare the frequency of letters in the message to how they normally appear in English.

I ran both methods on messages of different lengths to see how they perform. The idea was to learn which methods work better under different conditions and to understand what goes wrong when they fail.

2 How the Methods Work

Common Words Approach

This method is based on matching words. It tries every possible key (like a shift for Caesar or a pair of numbers for Affine), decrypts the text using that key, and then counts how many words in the decrypted message are found in a dictionary of common English words. The key with the most matches is picked as the best one. This approach is very direct—it works best when there are enough real English words to match.

Dot Product Approach

The dot product method is more mathematical. It creates a frequency vector (a list of how often each letter shows up) for the ciphertext and compares it to the known frequencies of letters in English using cosine similarity. The idea is that real English has a predictable pattern of letter usage (for example, 'e' is used a lot, 'z' not so much). If the decrypted text has a similar pattern, it's probably correct.

The problem is that in this experiment, half the words were gibberish—but the gibberish still followed English-like frequencies. That made it harder for the dot product approach to tell the difference between correct and incorrect decryptions, especially with shorter texts.

3 Results

This is the collected data. Each method was tested on 100 randomly generated texts for each word count.

Table 1: Decryption accuracy. Each method was tested on 100 randomly generated texts for word counts below 1000 and 20 trials for word counts above 1000.

Words	**Caesar Cipher**		**Affine Cipher**	
	Common Words	Dot Product	Common Words	Dot Product
10	63	21	13	33
50	82	34	54	38
100	100	68	84	59
1,000	100	75	100	75
10,000	100	75	100	75
100,000	100	75	100	75

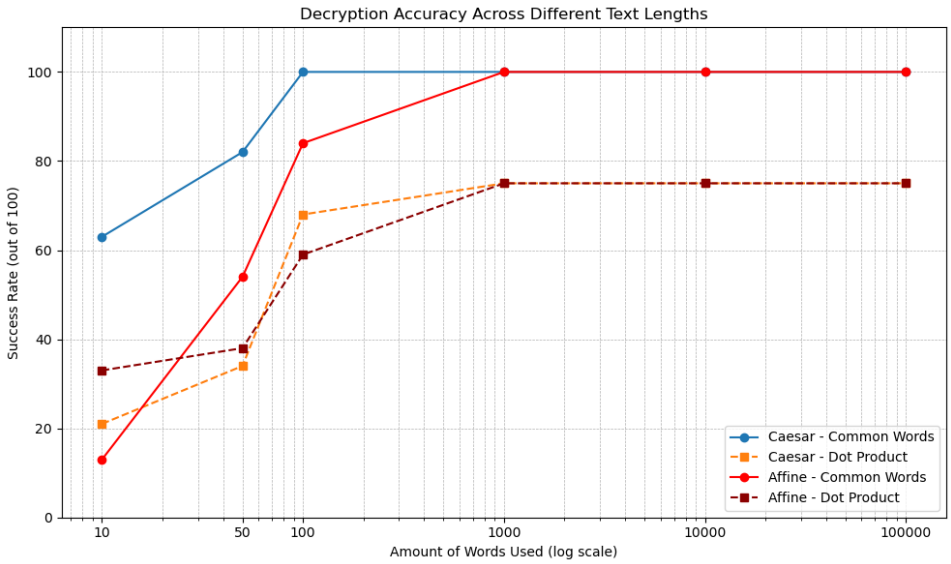

Figure 1: Decryption Accuracy Across Different Text Lengths

4 Analysis and Observations

The data clearly shows a few patterns:

- The **Common Words Approach** worked very well, especially for the Caesar Cipher. Even with just 50 words, it got over 80% accuracy. By 100 words, it had perfect results.
- The **Dot Product Approach** was much weaker for shorter texts. For example, with only 10 words, it only succeeded 21 times (for Caesar) and 33 times (for Affine) out of 100. It needs a lot of data to work well.
- With large enough text (1,000+ words), both methods performed well. But interestingly, the Dot Product Approach never got better than 75 successes out of 100. It hit a ceiling.

Why does this happen?

The Common Words Approach is good because it directly looks for actual English words. If it finds many, that's a strong sign the decryption is correct. The gibberish doesn't help it much, but it also doesn't fool it. This approach works well even with small messages.

The Dot Product Approach is more easily tricked. Since the gibberish follows English frequencies, it creates "fake" signals that confuse the method. The only time it does well is when there's enough real English in the message to overpower the noise from the gibberish. That's why it performs better as the text gets longer.

Other Languages: The key result of this experiment is that the Common Words Approach outperforms frequency-based methods when ciphertext contains deceptive patterns, such as gibberish designed to mimic English letter frequencies. This is important because many languages, such as Spanish and French, share broadly similar frequency distributions with English—for instance, vowels like "a" and "e" dominate across all three. A frequency-based method such as the Dot Product approach could therefore misidentify a Spanish or French message as English, or vice versa, since the statistical profiles are not distinct enough to guarantee accuracy. In contrast, the Common Words Approach relies on detecting real words rather than statistical similarity. As long as an appropriate dictionary exists for the target language, it will correctly distinguish between English, Spanish, French, or any other language. This demonstrates that dictionary-based methods are

more reliable for multilingual decryption, while frequency-based methods require caution whenever languages have overlapping statistical patterns.

5 Conclusion

1. The Common Words Approach is strong even on short messages. It's simple but very effective.

2. The Dot Product Approach only works well with long messages. It needs a lot of data to average out the noise.

3. When gibberish is made to match English frequencies, the Dot Product method struggles. It can't tell the difference between real and fake English-like messages unless the sample is big.

4. These results show that combining both methods might lead to better results—like using the Dot Product to narrow down the options, then checking them with the Common Words Approach.

5. This also tells us that if we want to use frequency-based methods on other languages, we need to adjust our frequency tables or models, or the results might be misleading.