Language Detection in Cryptography: Extending the isEnglish Method Across Languages Using the Common Character Approach

Shreyas Venkatesan, Eashaan Ranjith

July 2025

Abstract

Cryptography relies on encryption and decryption (with the key) being efficient, while cracking (decryption without the key) remains computationally hard. One important tool for cracking is an `IS-ENGLISH` program, which determines whether a given text is written in English. A common technique for such programs is the character frequency method, which evaluates text by comparing its letter frequencies to standard English distributions. This approach has historically been useful in frequency analysis attacks on classical ciphers such as the Caesar cipher. In this paper, we investigate whether this frequency-based method generalizes across languages by applying it to texts in German, Spanish, and Chinese. We also, as a playful experiment, test what happens when mismatched profiles are applied (for example, using English frequencies on Spanish text).

1 Introduction

Cryptography transforms readable text (plaintext) into an unreadable form (ciphertext), while decryption with the proper key restores the original message. Cracking, in contrast, refers to attempting decryption without access to the key. A common tool in cracking is an `IS-ENGLISH` test, which checks whether a candidate decryption is valid text. One classical approach to this is to analyze the distribution of letters in the candidate text and compare them to known frequencies in English. This method has proven effective in frequency analysis attacks on simple substitution ciphers such as the Caesar cipher.

In this paper, we expand upon the traditional isEnglish method by asking: Can this technique be generalized to detect multiple languages using the same frequency-based approach? We hypothesize that, since each language has a distinct letter frequency profile, it is possible to construct similar character frequency-based methods for other languages, such as Spanish, German, and Chinese. Furthermore, we predict that the precision of this classification will correlate with the uniqueness of the frequency distribution of a language.

To evaluate how well character frequency-based language verification generalizes beyond English, we conducted an experiment using texts in several languages that we encrypted with simple substitution ciphers (artificial rather than modern cryptographic systems). This ensured that frequency analysis techniques would be meaningful while keeping the encryption process transparent. For each case, the character frequency distribution of the text was ranked and compared to a reference profile for the language in which it was claimed to be written. This process used a modified version of the isLanguage method, which compares ranked character frequencies. The goal was not to identify the language of an unknown text, but to assess whether the frequency signature of a language is strong and distinct enough to verify the claimed language of a text.

Our findings indicate that the method performs reasonably well across multiple languages, particularly when the texts are sufficiently long to yield stable frequency patterns. For example, the method worked reliably with English and Spanish texts. In contrast, languages with more subtle or overlapping frequency profiles, such as French, posed greater challenges. These results suggest that frequency-based verification, while originally designed for English, can be extended to other languages with varying levels of accuracy.

2 isEnglish Method using the Common Character Approach

To explore statistical language detection in cryptographic analysis, we first developed a program that implements the isEnglish method using the common characters approach. This technique relies on the premise that letters in languages follow consistent frequency distributions. For example, in English, the letters 'E', 'T', 'A', and 'O' are significantly more common than letters like 'Q' or 'Z'. By modeling and comparing these distributions, we can statistically estimate the likelihood that a decrypted message belongs to a particular language.

2.1 Character Ranking Comparison Method

Instead of computing frequency vectors with exact proportions, our approach relies on the relative ranking of letters by frequency within a given text. For a cleaned input text T, we count the occurrences of each letter A-Z and sort them in descending order of frequency. This produces a ranked list:

$$\mathbf{R}_{\text{text}} = [r_1, r_2, \ldots, r_{26}]$$

where r_1 is the most frequent letter in T, r_2 is the second most frequent, and so on.

Each language L has a known standard ranking of letters based on typical usage patterns, denoted as:

$$\mathbf{R}_L = [s_1, s_2, \ldots, s_{26}]$$

To compare the text to its expected language profile, we compute a similarity score based on how closely the observed ranking $\mathbf{R}_{\text{text}}$ aligns with $\mathbf{R}_L$. One simple method is to count the number of top-k letters that overlap between the two rankings:

$$\text{Score}_L = |\{r_1, r_2, \ldots, r_k\} \cap \{s_1, s_2, \ldots, s_k\}|$$

A higher score indicates a stronger match between the text and the expected language profile. A threshold θ is selected such that:

$$\text{isLanguage}(T, L) = \begin{cases} \text{true}, & \text{if Score}_L \geq \theta \\ \text{false}, & \text{otherwise} \end{cases}$$

This rank-based method avoids reliance on precise frequency values and instead focuses on the relative ordering of character occurrences, making it simpler and potentially more robust across different languages and text lengths.

2.2 Results of the isEnglish Method

After implementing and testing the `isEnglish` method on a dataset of Caesar-encrypted English texts, we observed both expected and unexpected behaviors when determining whether decrypted outputs resembled English based on ranked letter frequency.

Expected Results

The method performed as intended under most controlled conditions:

- **Correct Decryption Identification:** When applied to English texts encrypted with a Caesar cipher, the method consistently identified the correct decryption key by selecting the version of the text whose ranked list of most frequent letters best matched the standard ranking of English letters.
- **Longer Text Performance:** For texts longer than approximately 100 characters, the `isEnglish` function showed high reliability. The larger text size provided a more stable and representative ranking of character frequencies, leading to consistent matches with the expected English ranking.
- **Low False Positives:** When run on random strings or non-English texts, the method rarely produced a close match to the standard English ranking, and correctly returned `false`, indicating the text was unlikely to be English.

Unexpected Results

Despite overall success, several limitations and anomalies were observed:

- **Short Text Inaccuracy:** For texts shorter than 50 characters, the observed character ranking was often too unstable to reliably match the English profile. In such cases, incorrect decryptions or non-English strings occasionally matched enough top-ranked letters to be falsely identified as English.
- **Coincidental Overlaps:** Some non-English texts produced letter rankings that happened to align with English (e.g., high frequency of `E`, `T`, `A`), resulting in false positives.
- **Unusual English Texts:** Texts composed of obscure English vocabulary, acronyms, or technical terms skewed the letter rankings away from typical English patterns, sometimes leading to false negatives.

These results highlight that the ranking-based `isEnglish` method is effective for general use but sensitive to input length and letter distribution noise.

3 How Is This Applied to Other Languages?

The original `isEnglish` method depends on the observation that some letters occur more frequently in English than others. To extend this idea, we tested whether comparing the **ranked list of most frequent letters** in a text to the **standard ranking for a given language** could work across multiple languages.

Standard Letter Rankings by Language

Each language has a characteristic ranking of letters based on typical usage. For example:

- English: `E, T, A, O, I, N, S, H, R, D`
- Spanish: `E, A, O, S, R, N, I, D, L, C`
- Chinese (Romanized Text): `A, I, O, N, E, G, H, R, Z, S`
- German: `E, N, I, S, R, A, T, D, H, U`

These standard rankings serve as reference profiles for each language.

Ranking Comparison Approach

Our multilingual adaptation proceeds as follows:

1. Count the occurrences of each letter (A-Z) in the input text.
2. Sort the letters in descending order by frequency to produce a ranked list.
3. Compare the top k letters in the text's ranking to the top k letters in the standard ranking for the target language.
4. Compute a similarity score by counting the number of overlapping letters in the top k positions.

For example, if 7 of the top 10 letters in the input text match the top 10 letters in the standard Spanish ranking, the similarity score would be 7.

Decision Rule for Language Verification

A threshold θ is defined for the minimum number of matching letters required to consider a text consistent with a given language. For instance:

$$\text{isLanguage}(T, L) = \begin{cases} \text{true}, & \text{if } |\text{Top}_k(T) \cap \text{Top}_k(L)| \geq \theta \\ \text{false}, & \text{otherwise} \end{cases}$$

This method does not attempt to identify the language from among many options but instead tests whether a text matches the **expected ranking** of a **known language**.

Practical Adaptations

To ensure reliable performance, several adjustments were made:

- **Normalization of Accented Characters:** Accented letters (e.g., `é`, `ñ`) were mapped to their unaccented equivalents (e.g., `e`, `n`) to align with standard rankings.
- **Text Length Threshold:** The comparison was only applied to texts with at least 100 characters, to reduce the impact of letter ranking instability.
- **Alphabet Consistency:** Only languages using the 26-letter Latin alphabet were included, allowing direct comparison across languages without requiring character set adjustments.

Results of Multilingual Application

After adapting the `isEnglish` method to handle other languages by comparing the frequencies of their most common letters, we tested the approach on encrypted and decrypted texts in English, Spanish, German, and Chinese (Pinyin). The experiment evaluated whether the method could correctly confirm the target language of a text after decryption. Overall, the method showed strong alignment with the expected frequency rankings of each language, while revealing several practical limitations.

Expected Results

When texts were sufficiently long (greater than 150 characters), the most frequent letters in the decrypted outputs closely matched the standard rankings for each language, making detection reliable:

- **English:** Texts consistently followed the expected order of `E`, `T`, `A`, and `O`. The prominence of `T` and `H` in particular made English distributions distinct from those of the Romance languages.
- **Spanish:** Results reflected the ranking `E`, `A`, `O`, and `S`. The especially high counts of `A` and `O` clearly separated Spanish texts from English and German, which rely more on `T` and `N`.
- **German:** Decrypted German texts displayed strong frequencies of `E`, `N`, `I`, and `S`, with `U` ranking noticeably higher than in English or Spanish. This alignment with the German profile enabled consistent detection.
- **Chinese (Pinyin):** Recognition worked well due to the heavy use of vowels (`A`, `I`, `O`) and the distinct presence of `N`, `G`, and `Z`, which created a profile unlike the European languages.
- **Decryption alignment:** When paired with the Caesar cipher process, the method often confirmed the correct change by showing that the decrypted text letter frequencies fall into the expected ranking for the target language.

Cross-Language Comparisons

Although the method was not primarily designed to compare languages against each other, we tested whether texts might be misclassified when analyzed against multiple profiles:

- **Correct separation at high thresholds:** For longer texts and stricter similarity measures, each language consistently detected only itself. Even when distributions appeared superficially close (e.g., Spanish and Pinyin both having high vowel counts), the finer ordering of letters such as `S`, `R`, and `G` provided separation.

- **Apparent similarities at low thresholds:** For shorter texts (under 100 characters), fluctuations blurred distinctions. For example, Pinyin samples lacking `Z` or `G` sometimes looked closer to Spanish, and English with unusually high `A` or `O` usage occasionally resembled Spanish distributions.

Unexpected Results

Some factors outside of raw length also impacted stability:

- **Accented characters:** In Spanish and German, accented letters were normalized to their unaccented equivalents (e.g., `á` → `a`, `ü` → `u`). This often inflated already frequent vowels, making Spanish texts appear more vowel-heavy and pushing German `U` counts higher than expected. While detection still succeeded on long samples, this occasionally shifted results for shorter ones.
- **Stylistic variation:** Texts dominated by proper nouns or specialized terminology disrupted typical frequency patterns, reducing accuracy when the vocabulary did not reflect the language's standard distribution.
- **Mixed-language samples:** In bilingual outputs, the frequency curves flattened into hybrids (e.g., a mix of Spanish vowels with German consonants), producing near-equal matches across two profiles.

Just-for-Fun Explorations

While applying an `IS-ENGLISH` test to Spanish text is not a rigorous cryptanalytic method, we include it here as a playful exploration. It highlights how mismatched frequency assumptions can yield misleading results, and illustrates the limitations of the common-letter approach.

Summary of Findings

Across all tests, the adapted method successfully identified English, Spanish, German, and Chinese (Pinyin) texts when samples were long enough to reflect their characteristic letter rankings. While shorter samples, accent normalization, and mixed-language inputs introduced variability, higher thresholds ensured that each language reliably recognized itself and not another. These results suggest that letter-frequency profiling remains a strong tool for language verification in cryptographic contexts, provided its known limitations are taken into account.