Pre-Analysis

The origin of the text

In order for the effort of decoding the runes to be useful, it helps to know what to look for. A big question that many have had for both my solution and the text as a whole is "How do we know it's Romaji?". There are a few ways:

1. Index of Coincidence Scoring

We can look at the index of coincidence[1] (IOC) of the ciphertext. In short, it's a method of determining not only if text is random gibberish or real, but it can also be used as an indicator of the language. For the sake of this explanation, the IOC scores are not normalized. to get a baseline for random text for a language, you can use this formula:

- `1 / N` where N is the size of the alphabet. For many latin based languages, N = 26. For Romaji, N = 21 (though Romaji is represented with a latin alphabet, a number of letters aren't used). For English and Romaji, you get the following two values.
    - English Random Text: 0.038
    - Romaji Random Text: 0.047
- To get scores for non-random text in both languages, you can use the algorithm found in the Index of Coincidence wikipedia article[1]. For convenience, the non-normalized scores for English and Romaji are included. You can also calculate these scores for ciphertext using the Index of Coincidence tool at a site like dcode.fr[2]
    - English Non-Random Text: ~0.06
    - Romaji Non-Random Text: ~0.089

To give an example of how reliable this method is for providing an indicator* of the source language, lets take ~40 random characters of text from the Sheikah tapestry. It's an excellent example because it's provided in both languages. The text*:

`The royal family fear and exile the Sheikah tribe's` and `Ōke wa Sheikah-zoku no chikara o osore, tsuihō su` calculate to the following scores:

- **English**: 0.063

- **Romaji**: 0.074

You're probably asking why the Romaji isn't 0.089 or above. The reason is that English words are mixed in, and through the normalization process of scoring, it lowers the score to be mostly between English and Romaji. This method is incredibly useful because it works, even with a limited set of known alphabet characters. All this has been a primer for measuring and explaining the theory that the ciphertext is Romaji. The ciphertext `YDEBSXDSAZYDBZRCESRABCDEGCGCDEXABABCDECD` IOC score is **0.089**.

2. Frequency as an indicator for source language

Another way that we can get an indicator of source language is by observing the delta of frequency percentage of the letters in the ciphertext. For the sake of using a known text source to help explain, let's first compare the frequency curve of the sample English Sheikah tapestry text with the Romaji.

English

```
The royal family fear and exile the Sheikah tribe's
```

Romaji

```
Ōke wa Sheikah-zoku no chikara o osore, tsuihō su
```

Using a frequency analysis tool like the one at Boxentriq[3], you can observe the following information for single character frequency distribution:

- **English**: The letter with the highest frequency, E is 5% higher than the next letter. The curve then becomes steady with a 1-2% delta between letters before becoming steep and less evenly distributed

- **Romaji**: The letter with the highest frequency has a delta between itself and the next letter of ~2.6% with a steady decrease of ~1% between letters. The curve for Romaji text is shallow and the distribution is more steady.

- **Ciphertext**: Looking at the distribution for the ciphertext `YDEBSXDSAZYDBZRCESRABCDEGCGCDEXABABCDECD`, we can see it follows more closely with the Romaji analysis.

Note: It's also worth mentioning here that if you look at the frequency percentage in both texts, Romaji has more characters with the same percentage for scenarios where the number of instances of the character is > 1. This is because most of Romaji text ends with a vowel and the number of vowels are limited. English text has more characters with single uses because it has different consonant formation rules. English even has entire words that are made up of nothing but consonants.

Footnotes

- Indicator is used because it's possible to skew these scores in a number of ways, but typically it's pretty obvious when that occurs.
- The scoring of the Sheikah tapestry text is done without any spaces or diacritics to stay consistent.

Summary

The ciphertext has an IOC and frequency distribution that is in line with Romaji text of the same sample size. The probability of the ciphertext being Romaji is high.

Links

1. Index of Coincidence, https://en.wikipedia.org/wiki/Index_of_coincidence
2. Calculate Index of Coincidence, https://www.dcode.fr/index-coincidence
3. Boxentriq Frequency Analysis Tool, https://www.boxentriq.com/code-breaking/frequency-analysis