**Estimates on Letter and Character Frequency in a CW Contact (QSO)**

This is prepared for generation of data for machine learning of Morse Code for decoding CW using a skimmer or using Basian probabilities to improve CW or perhaps RTTY [decoders using other methods](http://fweb.wallawalla.edu/~frohro//CW-Robot/Matched%20Filter%20MAP%20Detector%20Paper.pdf). The letter frequencies for the general English language can be found online easily, however, I was unable to find them for the specific area of a CW QSO (ham radio conversation using the CW mode). I am therefore attempting to approximate them using my experience and other data I can find online.

First start with the [letter frequency in English](http://cs.wellesley.edu/~fturbak/codman/letterfreq.html).

Now, [the average length of a word in English is 4.5 letters](https://www.researchgate.net/publication/230764201_Average_word_length_dynamics_as_indicator_of_cultural_changes_in_society). A space appears after every word. Furthermore in CW conversations, whenever an operator is thinking repeated extra spaces can occur. If we ignore these the space is the character about 1/(4.5)\*100% of the time.

In regular English text, [punctuation occurs with this frequency](https://www3.nd.edu/~busiforc/handouts/cryptography/letterfrequencies.html):



In ham radio QSO’s in my experience, this is quite different. For example, most hams only know the Morse code for period, comma, question mark, and slash. The slash is used more than in general English, as is the question mark, because hams send this to indicate they missed something, which happens fairly often. Commas are not used as often as in regular English either. Periods are probably used less too, as they take a long time to send compared to letters, and CW is more informal than regular English text. Based on these assumptions I am going to guess frequencies for this punctuation per 1000 words as:

|  |  |
| --- | --- |
| **Punctuation** | **Estimated Frequency per 1000 Words** |
| Period | 50 |
| Comma | 40 |
| Question Mark | 20 |
| Slash | 3 |

Furthermore, they use prosigns, punctuation that are most often recognized by hams as two letters run together, often written with an overscore over both characters, and more recently mapped to other unused characters in the Morse code mapping, like + for $\overline{AR}$ (end of transmission) or = for $\overline{BT}$.

Then there are the numerical digits. These occur more often than in regular text, because every call sign has at least one of them, and the call sign of both stations is usually sent at the beginning and end of every transmission, and because CW is similar to texting, it is unusual for a number to be spelled out; furthermore, signal reports are numbers, and they are usually sent for each QSO. I will assume no digit is more likely than another. I’m going to guess a digit occur four times every transmission, and estimate each transmission is 25 words, so that is four times as often as $\overline{AR}$ or 16 times per 1000 words for each digit.

Here is a list of more commonly used [prosigns](https://en.wikipedia.org/wiki/Prosigns_for_Morse_code) in my experience, with my estimates of their frequency per 1000 words.

|  |  |  |
| --- | --- | --- |
| **Prosign** | **Meaning** | **Frequency (per 1000 words)** |
| $$\overline{AR}$$ | used after every transmission, just before turning it over to the other operator for his transmission  | 40 |
| $$\overline{BT}$$ | signifying the end of that paragraph, often used when the operator is thinking | 2 |
| $$\overline{AS}$$ | please wait just a moment | 1 |
| $$\overline{HH}$$ | oops, I made an error sending that just now, let me try again | 2 |
| $$\overline{SK}$$ | sent at the end of a contact (QSO) | 2 |

To obtain the final probability estimates, we need to put this all together. I will assume there are 1000 words, and letter frequencies are $1000P\left(letter\right)4.5{letters}/{word}$. We renormalize, based on how many characters we have, and recompute probabilities using $P\left(chararter\right)=frequencyper1000{words}/{characters}per1000words$

When generating the characters for training in a machine learning algorithm, use a uniformly distributed random number generator, and partition the interval (0, 1) into lengths of the probability we compute in the table below. When the number occurs within the interval, we generate the character that goes with that interval.

The total number of characters per 1000 words is:

$$ncptw=1000words\left(\left(4.5+1\right){letters}/{word}\right)+\left(50+40+20+3+160+40+2+2+1+2\right)characters=5820characters$$

For those characters we have the frequency per 1000 words, the probability is:

$$P\left(character\right)=characterfrequency\in 1000{words}/{ncptw}$$

For the letters the probability is:

$$P\left(character\right)=P{\left(letter\right)∗4.5∗1000}/{ncptw}$$

For the space:

$$P\left(space\right)=1000{\left({spaces}/{1000}words\right)}/{ncptw}$$

These are used to compute the following table.



These values can be recalculated using the Letter Probabilities spreadsheet. The probability of each character is given by the P(character) column, and the acpw stands for “Average Characters per Thousand Words”, and ncptw stands for “Net Characters per Thousand Words.” The column before the P(character) is the estimate of number of characters per thousand words for the symbol. The value after the ncptw is the total probability as a check.