Project 2: Approximate Counting

Tiago Fernandes, 88784

***Abstract* – In this project, one exact counter and two approximate counters (one with fixed probability and another with decreasing probability) were studied. It was possible to count the letters of three book editions with the three counters, and determine the most common and least common letters. The exact counter gave the most fine-grained results, but was most memory-heavy. The counter with fixed probability allowed to successfully find the most common letters in order, but was less reliable for the least common letters, while using less memory. On the other hand, despite it's significant memory reduction, the Morris counter allowed to find some of the most common and least common letters, but showed huge relative errors in some cases.**

***Keywords* – Counting, Approximate counting.**

I. INTRODUCTION

The counting process was developed long before the time of recorded history, with archaeological evidence that counting was employed by humans for at least 50,000 years [1]. With the gradual evolution of society, simple counting became imperative, for instance to keep track of social and economical data of a tribe.

Nowadays, as a consequence of the digital revolution, there is more and more data being created every day: just in the year 2020 there was 59 ZB (59×10^{21} bytes) of data created, captured, copied and consumed worldwide [2]. This huge amount of data generated daily has the potential to be analysed, allowing to discover useful information and support decision-making. However, the volume of the data is also a challenge, as the classical methods would require too much hardware resources and energy. Thus, probabilistic methods like approximate counting are used in order to obtain approximate answers using less resources.

In this project, the goal is to count the number of occurrences of letters in text files, using both exact and probabilistic counters, and to compare the efficiency and limitations of the different counters.

II. COUNTING ALGORITHMS

Three different counters were used in this project: an exact counter, an approximate counter with fixed probability and another approximate counter but with decreasing probability. The three mentioned counters were implemented in Python.

The exact counter is the simplest possible counter. For each new item, it increments the counter. It requires n bits to be able to count up to $2^n - 1$ events. This is the best that can be done deterministically, but in some cases this memory requirement is too expensive.

A simple alternative would be to count every other new item with a fixed probability p, as shown in Alg. 1. If $p = 0.50$, this can be imagined as tossing a coin for each new item and only incrementing the counter if the coin lands heads.

Algorithm 1 Probabilistic counter with fixed probability

```
Require: $p \in [0, 1]$
  S ← 0
  for each item do
      r ← Uniform(0,1)
      if $r < p$ then
          S ← S + 1
```

In the general case, when counting n events this probabilistic counting method would yield an expected value of $n \times p$, with a standard deviation of $n \times p \times (1 - p)$. Thus, the count estimation is obtained dividing the counter value by p. Assuming p has the form of $1/2^k$, it would allow to count up to 2^{n+k} events using n bits, instead of the $n + k$ required if an exact counter was used.

This algorithm has two main drawbacks, the first being that small numbers are counted with high relative errors. For example, if $p = 1/32$ and there are 64 items, there is a probability of more than 10% ($\Pr(0; 64, 1/32) \simeq 0.13$) that the counter does not increment a single time and yields an estimate of 0. Moreover, there is also the problem that for very big numbers, it would be possible to be more economical by using even smaller probabilities.

In 1978, Robert Morris devised a simple counter that tackles these two drawbacks by using a variable probability, which is higher when the counter has lower values, in order to count low numbers with less errors, and lower as the counter value increases, in order to be more memory efficient. The pseudocode of this probabilistic counting algorithm, known as Morris counter, is shown in Alg. 2.

Algorithm 2 Probabilistic counter with decreasing probability

```
Require: $p \in [0, 1]$
  S ← 0
  for each item do
      r ← Uniform(0,1)
      if $r < p^S$ then
          S ← S + 1
```

It can be shown [3] that if k is the value of a Morris counter (with initial probability $p = 1/a$) after n increments, then $(a^k - a)/(a - 1)$ is a good estimator of n. Solving in order of k, we get the expected counter value: $\log_a[n(a-1) + a]$. Assuming $p = 1/2$, this counter allows to count up to 2^{2^n}, using just n bits, which is much better than the previous counters.

III. Text Preparation

In order to test the implemented counter, three versions of the 1865's Jules Verne novel *From the Earth to the Moon: A Direct Route in 97 Hours, 20 Minutes* were chosen: the original French book, and both the English and Portuguese translations, obtained from the Project Gutenberg [4]-[6].

Each plain text file was downloaded and processed as follows in order to transform the plain text books into a long string of capital letters. First, the Project Gutenberg's headers and license information were removed, in addition to the translation notes for the English and Portuguese editions. In the case of the English edition, a sequel novel that was also included in the file was removed too. Then, all non-alphabetic characters were removed from each file and the accented letters were converted to the corresponding regular letters. Finally, all letters were capitalized and the resulting strings were saved.

IV. Results and Discussion

A. *Exact counter*

The exact counter was only run once for each file, as the results would always be the same due to its deterministic nature. This counter gives an exact description of the text samples, which will be useful later to access the information obtained by the approximate counters. Additionally, these results allow to determine if the chosen text samples can be regarded as typical examples of the respective languages and if there are some particularities that might affect the results.

Before analyzing the occurrences of each letter, the total number of letters was found for each book edition by summing the values of the 26 different counters (one for each letter). It was found that the English edition was the smallest, with 235645 letters, followed by the French and Portuguese versions, with 266484 and 278249 letters, respectively. The French and Portuguese have approximately the same size, but the English version has about 12% less letters than the original book in French, which is a significant difference.

After ensuring that there was no problem with the English version source file and the text preprocessing pipeline, it was concluded that the size differences between the languages were not an error. In fact, there were found multiple accounts (e.g. [7], [8]) of the different text sizes in translation which pointed out that the English language is more compact than Romance languages like Portuguese and French, as observed in this project. For this reason, it was decided that the results were better presented as relative frequencies instead of absolute values.

The counter values for each letter were then analyzed. Initially, the relative frequencies of each letter for the three languages were plotted, as shown in Fig. 1. It can be seen that the letter frequencies are, as expected, different for each language. For instance, 'A' is the most common letter in Portuguese, with a relative frequency 5 percentage points higher than the other languages. In the case of both French and English, 'E' is the most common letter, albeit with a significantly higher abundance in French. 'E' is also the second most frequent letter in Portuguese, with a similar frequency as English. Letters like 'O', 'I' and 'T' are also popular in all languages. On the other side of the spectrum, letters such as 'Y' and 'W' are almost non-existent in Portuguese and French, but are present (although still rare) in English. Conversely, 'Q' appears about 1% of the time in the two Romance languages, it appears much less in English. Other letters such as 'J', 'X', 'K' and 'Z' are very rare across all languages.

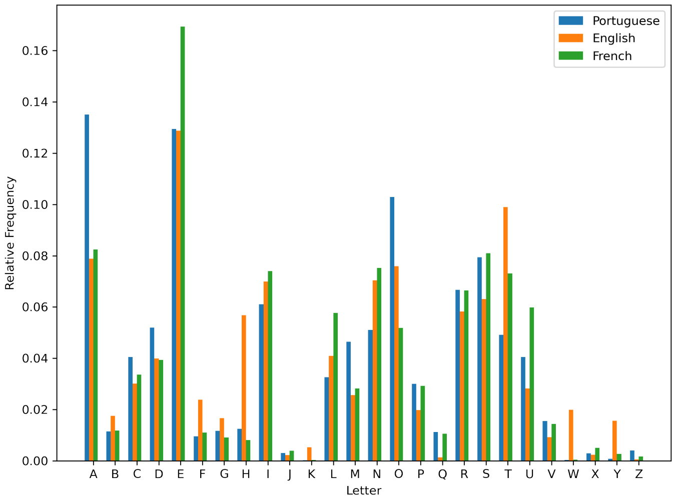

Fig. 1: Comparison of the letter frequencies in the different languages.

The results for each respective language were analyzed further. In Fig. 2, the measured frequencies are compared with the relative frequencies of a typical text in that language, obtained from [9]. Overall, every result is in accordance with the expected values, with all deviations below one percentage point. This shows that the chosen text samples are representative of the typical text in the respective language. The plots also allow to visually identify the most common and least common letters in each language.

In the Portuguese text, 'A', 'E', 'O', 'S' and 'R' are, in that order, the five most common letters. The five least common letters are, from least to most common, 'K', 'W', 'Y', 'X', 'J'. Comparing to the expected results, the most common letters are all in the predicted order, although 'A' appears only 13.5% of the times, instead of the expected 14.6%. This could be due to some error while converting accented characters like 'ã', 'á', or 'â' to 'A', but no such case was found. Another possible explanation is related to the fact that the text was written in 1874, and the relative frequencies of the letters has slightly changed since then.

Regarding the five least common letters, they are the ones expected but in a slightly different order. 'K', which was expected to be the third least common letter, with frequency of 0.02%, was the least common, but with a higher frequency, 0.03%. Despite this exception, the other letters were ordered as expected, although 'W', 'Y' and 'X' appeared more times than predicted. This can partially be explained by the fact that part the plot of the book takes place

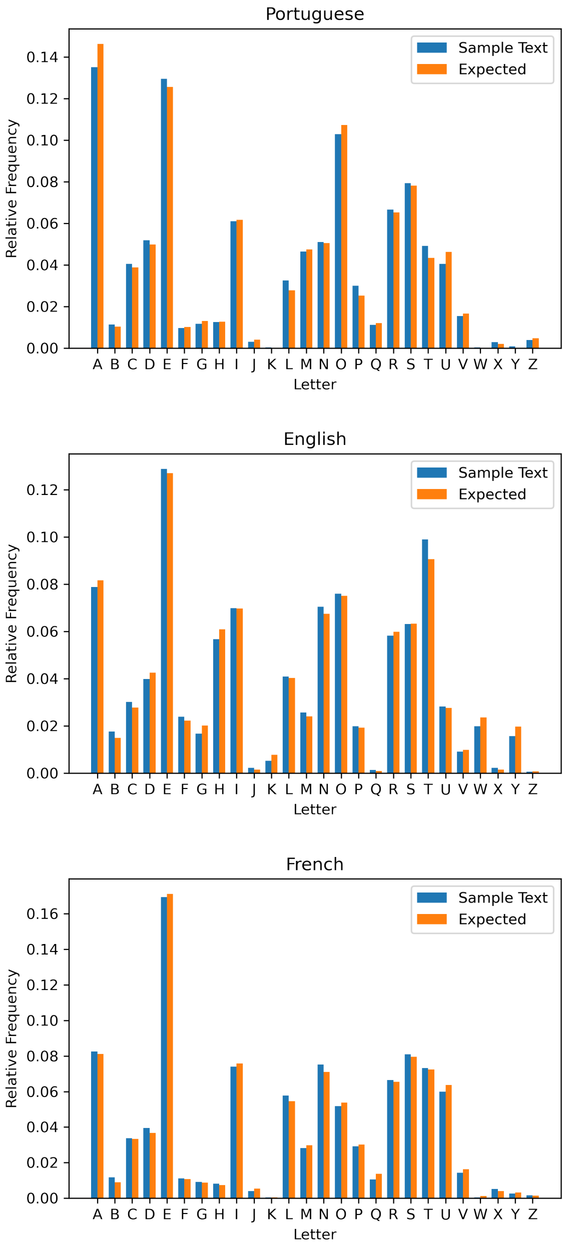

Fig. 2: Comparison between the relative frequencies in the three different book editions and the frequencies reported in [9].

in the United States, so cities like ’New York’ and ’Washington’ and words like ’whisky’ and ’sandwich’ increase the number of ’W’s and ’Y’s. As to the number of ’X’s, the 51 occurrences of forms of the verb ’exclamar’ can help to explain the over-representation. It should be noted that ’K’ and ’W’ only appeared 73 and 100 times, respectively, so it is expected that at least the first approximate counter can have difficulties in these letters for the Portuguese edition.

As to the English text, the five most common letters were, in order, ’E’, ’T’, ’A’, ’O’ and ’N’. ’T’ appeared more than expected, with a frequency of 9.9% instead of the expected 9.0%. Besides, ’N’ also appeared a little more than expected, making it slightly more frequent than ’I’, which should be the fifth most common letter. The least common letters were, in order, ’Z’, ’Q’, ’J’, ’X’ and ’K’, closely following predictions. It should be noted that in English only one letter, ’Z’, appears less than 0.1% of the time, with only 124 occurrences. The other letters shouldn’t pose much problems to the fixed probability counter, as ’Q’ already appears more than 300 times.

Finally, the French language seems to be the one were the results most closely follow the expected values, which could be related to the fact that the French text is the original book version, and the other two are translations which can be biased by the words and sentence structures of the French. The most common letters are ’E’, ’A’, ’S’, ’N, and ’I’, and the least common are ’K’, ’W’, ’Z’, ’Y’ and ’J’. Both ’K’ and ’W’ could be hard to count, as they have 81 and 104 appearances, respectively.

B. *Approximate counter with fixed probability*

An approximate counter with $p = 1/16$ was tested in the generated text samples, by running it a few times for each file, and the results were saved. Fig. 3 compares some distributions obtained for 100 repetitions with the respective distributions for 1000 repetitions. With 1000 repetitions, the results resemble a normal distribution, even for the rare letter ’K’ which shows a slightly anti-symmetrical distribution due to the accumulation of values at 0. Thus, 1000 runs per file was a good compromise, as letter counts distributions very close to a Gaussian can be obtained while not taking too long. Only the results for 1000 repetitions were used in the rest of the analysis.

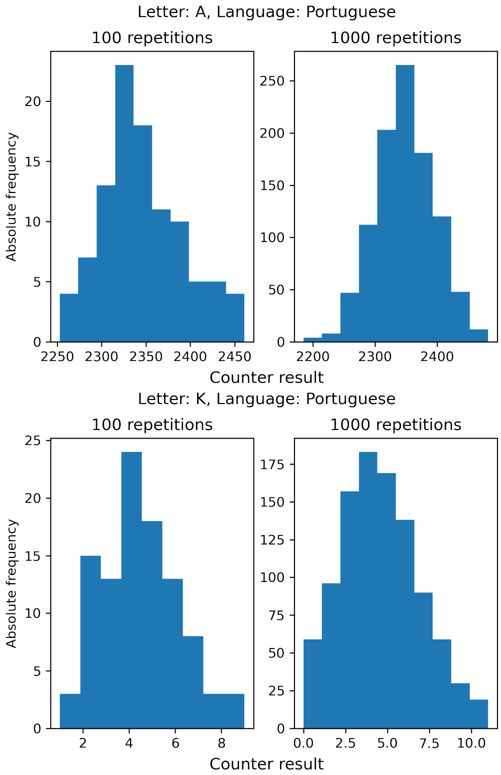

Fig. 3: Distributions of the fixed probability counter results for the most and least frequent letters of the Portuguese text.

Initially, the mean value of each counter was compared with the theoretical expected value presented in section II, and excellent agreement was found in all cases. The results for the Portuguese text, plotted from highest to lowest average, are shown in Fig. 4. The results for the other languages are shown in Appendix A.

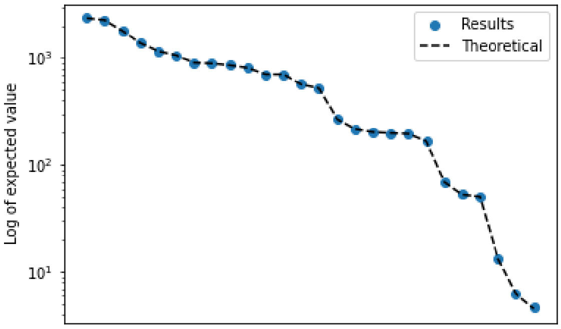

Fig. 4: Comparison between the mean value of each letter counter and the theoretical expected value, for the Portuguese text sample using the fixed probability counter.

Then, the counter estimations were obtained by dividing the counter values by the probability of incrementing the counter, which is the same as multiplying by 16. A random repetition of the counting was chosen and compared with the values obtained by the exact counter. Fig. 5 shows the comparison for the Portuguese text, in terms of the relative frequency of each letter, while the results for the other languages are shown in Appendix B. In all cases the counter approximated the real letter frequencies fairly well, although it showed, as expected, more difficulties for the rarest letters.

In the shown graph, all the five most frequent letters were identified in order, with relative errors below 5%. As to the less frequent letters, only 'W' and 'K' were swapped, since the 'W' count was underestimated, with a relative error of 36%. Relative errors above 10% were also observed for the other rare letters. In the other languages, the fixed probability counter also allowed to identify the most common letters with relative errors below 5%, only with occasional difficulties in letters with very similar real relative frequencies. The less frequent letters could all be identified in order using the chosen counter repetition, but the high relative errors over 30% mean that other repetitions would probably not yield correct results, as was verified.

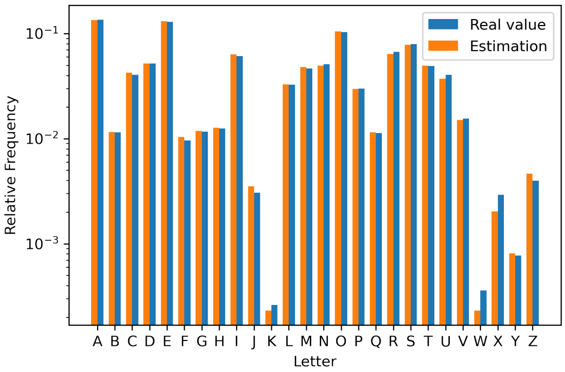

Fig. 5: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the fixed probability counter, for the Portuguese text. The results are shown in logarithmic scale to better visualize lower frequencies.

Finally, a table with the results for the letter counters of the five most and least frequent letter, using the data of the 1000 repetitions over the Portuguese text sample, is presented in Table I. The second column contains the real absolute frequencies of the letters in the text, obtained by the exact counter, while the third column shows the average value of the estimations yielded by each counter, which ideally should be equal to the previous column. The minimum and maximum values for each counter and the standard deviation are also shown, to better evaluate the variability of each counter results. Finally, the second to last column shows the relative error between the average estimation and the exact counts, and the last column shows the relative error of the worst estimation obtained for each letter. The full table as well as the tables for the other two languages are shown in Appendix C.

The average estimations were very close to the exact count value, as expected after the analysis of the graph of Fig. 4. Over all languages, the biggest relative error was 2.6%, ob-

TABLE I: Results of the approximate counter with fixed probability, ran 1000 times in the Portuguese text. Only the results of the five letters with the highest and lowest estimation values are shown.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
A	37584	37569.84	34960	39712	757.10	0.04	6.98
E	36029	36029.97	33296	38416	754.47	0.00	7.59
O	28623	28642.32	26592	31520	643.99	0.07	10.12
S	22100	22098.14	19680	24096	585.67	0.01	10.95
R	18555	18551.57	16768	20656	538.20	0.02	11.32
J	849	851.34	528	1216	112.09	0.27	43.23
X	813	803.98	496	1168	113.60	1.11	43.67
Y	214	213.42	64	416	54.90	0.27	94.39
W	100	99.42	0	240	38.68	0.58	140.00
K	73	74.94	0	176	34.64	2.66	141.10

tained for the 'K' in the Portuguese edition, so even more repetitions would be needed to get a better result. However, almost all the other relative errors were below 1%. It is observed that the standard deviation grows from the bottom to the top of the table, but the average value grows much more, which translates into much more precise estimates for the most common letters. It was also found that there were cases were the counters of rarest letters were never updated and others were the estimated value was more than double the original, yielding relative errors higher than 140%. Generally, the relative error between the worst estimate and the exact count increases as the frequency of the letter decreases, which confirms the difficulty of counting the least common letters.

C. *Approximate counter with decreasing probability*

A Morris counter with initial probability $p = 1/\sqrt{3}$ was tested in the same way as the previous counter, for the three different files and with a different number of repetitions. Since the expected values were smaller than the ones of the previous counter, the typical distributions for 10k repetitions, in addition to 100 and 1k repetitions, were also visualized; those are shown in Fig. 6.

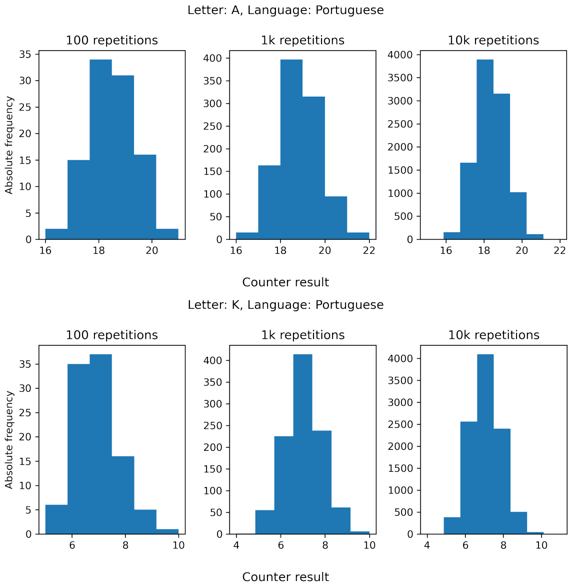

Fig. 6: Distributions of the Morris counter results for the most and least frequent letters of the Portuguese text.

There seems to be no improvement from 1k to 10k repetitions in terms of the shape of the distribution, so the number of repetitions was kept at 1000. It is clear that the Morris counter yields low values for even the most common letters, as the counter value for the letter 'A' was about 100 times lower than the previous counter. It is also observed that there are no cases of null counter values, and even for the least frequent letter the smallest count is 5. Moreover, the distributions also seem to be narrower than when using a fixed probability counter.

Once again, the mean value of each counter was compared with the expectation value, calculated as indicated in section II. The results for the Portuguese text are shown in Fig. 7. There seems to be good agreement, but there is an approximate constant difference around 0.25 between the two values for all counters. This behaviour also happened for the other text samples, as shown in Appendix D. The origin of this deviation could not be determined, but the slight bias towards lower values than expected should be taken in account henceforth.

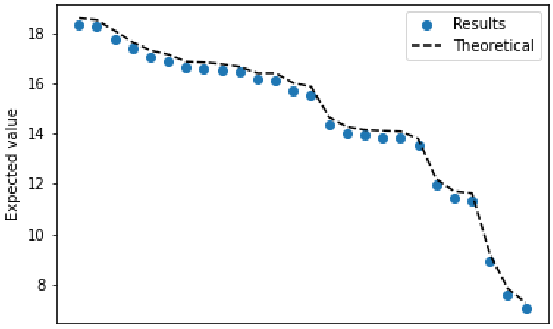

Fig. 7: Comparison between the mean value of each letter counter and the theoretical expected value, for the Portuguese text sample using the Morris counter.

The counter estimations were obtained by computing $(\sqrt{3}^k - \sqrt{3})/(\sqrt{3}-1)$ for every counter. The counter results of a single random repetition were chosen and compared with those of the exact counter in terms of relative frequencies, as shown in Fig. 8 for the Portuguese text sample and in Appendix E for the other languages.

It can be seen that there are many letters with the same estimated relative frequency, which happens because the counter counts the logarithm of the real value, so only powers of $\sqrt{3}$ (to be precise, a linear transformation of powers of $\sqrt{3}$) can be estimated. This problem is particularly inconvenient when trying to order the most common letters, for instance. In the run of the counter shown, the value of the letters 'A', 'E', 'O' and 'S' is the same, which transpires to estimations with relative errors up to 110% and doesn't allow to order the most common letters. These effects were also observed for the other languages, where not all the top five letters could be identified and relative errors as big as 200% were present. (A particularly extreme example was found in the case of French, where the second most common letter 'A' was estimated to be the tenth most common.) For the less frequent letters, the four least frequent letters in Portuguese were identified almost in the correct order and even better results were obtained for the other languages.

Finally, a table similar to the one shown for the fixed probability counter is shown in Table II, and the full table and the tables for the other languages are shown in Appendix F. All the average estimation values are lower than the real value, with relative errors around 12%. Despite this bias already expected from the analysis of Fig. 7, the average estimation sorts the letters in their frequency order. However, the Morris estimation shows deviations much higher than the previous estimator, since the standard deviations are the same order of magnitude as the average value for

TABLE II: Results of the approximate counter with fixed probability, ran 1000 times in the Portuguese text. Only the results of the five letters with the highest and lowest estimation values are shown.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
A	37584	32782.28	8960.13	241984.94	24905.30	12.78	543.85
E	36029	31424.52	8960.13	241984.94	22122.90	12.78	571.64
O	28623	23720.01	5172.13	139709.07	15860.34	17.13	388.10
S	22100	19420.90	5172.13	80660.07	12420.39	12.12	264.98
R	18555	15892.16	2985.13	80660.07	10420.62	14.35	334.71
J	849	731.78	189.28	5172.13	497.26	13.81	509.20
X	813	687.22	108.28	2985.13	456.66	15.47	267.17
Y	214	180.94	34.52	993.47	120.56	15.45	364.24
W	100	84.76	18.93	993.47	64.89	15.24	893.47
K	73	62.94	9.93	329.58	45.48	13.79	351.48

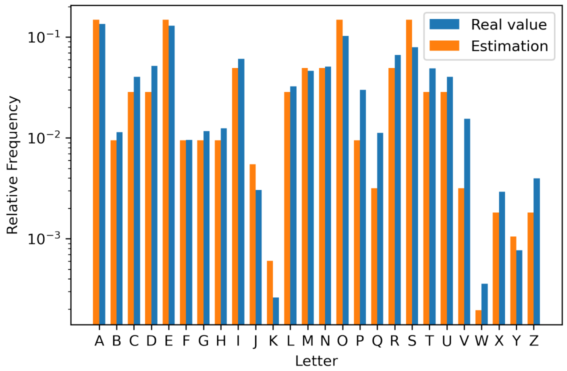

Fig. 8: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the Morris counter, for the Portuguese text. The results are shown in logarithmic scale to better visualize lower frequencies.

all letters. This transpires in the huge relative errors of the worst estimations, with some predictions being more than 5 times above or below the real value.

V. CONCLUSIONS

In this project, the efficiency and limitations of one exact counter and two approximate counters were studied by counting the number of letters in a book and two translations of the same book, and trying to identify the most common and least common letters.

The approximate counter with fixed probability was found to be very precise in the identification of the most used letters, but for the least frequent letters it could not always identify them in order. The absolute frequencies of the letters could be estimated with errors around 10% for the most common letters and errors sometimes over 100% for the least common.

The approximate counter with fixed probability showed a constant bias in its estimations, predicting on average absolute frequencies around 12% below the real value. Moreover, using this counter it was almost impossible to find out the order of the most common letters, which would end up with the same estimation due to the low number of possible estimation values for higher numbers, but it allowed to identify the least common letters. Nevertheless, this counter requires much less memory than the others, as the maximum value it reached was 21 (5 bits), compared with values in the order of thousands for the other counters (12 and 16 bits for the fixed probability and exact counters, respectively).

However, only the exact counter allowed to compare the letter frequencies of the different language editions, and also allowed to find some small deviations between the frequencies in the texts and the average frequencies of the respective languages, which would not be possible to find using the implemented approximate counters. Despite not always being an ideal (or possible) choice for all applications, it proved to be the most suitable counter for this specific case, due to its intrinsic correctness and its simplicity.

REFERENCES

[1] Howard Eves, *An Introduction to the History of Mathematics*, Saunders College Publishing, 6 edition, 1990.

[2] "The Conversation: The world's data explained", https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964, Accessed: 30/12/2021.

[3] Philippe Flajolet, "Approximate counting: A detailed analysis", *Bit*, vol. 25, pp. 113, 1985.

[4] "Project Gutenberg: The Moon-Voyage by Jules Verne", https://www.gutenberg.org/ebooks/12901, Accessed: 28/12/2021.

[5] "Project Gutenberg: De la Terre à la Lune by Jules Verne", https://www.gutenberg.org/ebooks/799, Accessed: 28/12/2021.

[6] "Project Gutenberg: Da terra à lua, viagem directa em 97 horas e 20 minutos by Jules Verne", https://www.gutenberg.org/ebooks/28341, Accessed: 28/12/2021.

[7] "Stack Exchange: Do most languages need more space than English?", https://english.stackexchange.com/a/3022, Accessed: 02/01/2022.

[8] "Why does content expand when translated?", https://www.inter-contact.de/en/blog/text-length-languages, Accessed: 02/01/2022.

[9] "Wikipedia: Frequência de letras", https://pt.wikipedia.org/wiki/Frequ%C3%AAncia_de_letras, Accessed: 28/12/2021.

APPENDIX A - EXPECTED VALUE OF THE APPROXIMATE COUNTERS WITH FIXED PROBABILITY

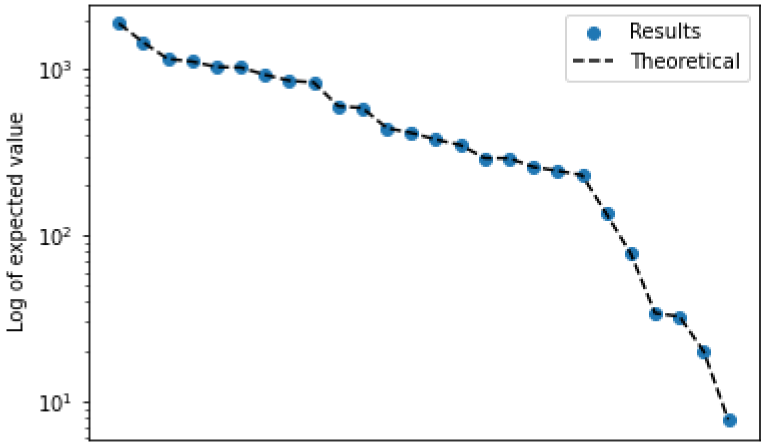

Fig. 9: Comparison between the mean value of each letter counter and the theoretical expected value, for the English text sample. Points were ordered from highest to lowest mean.

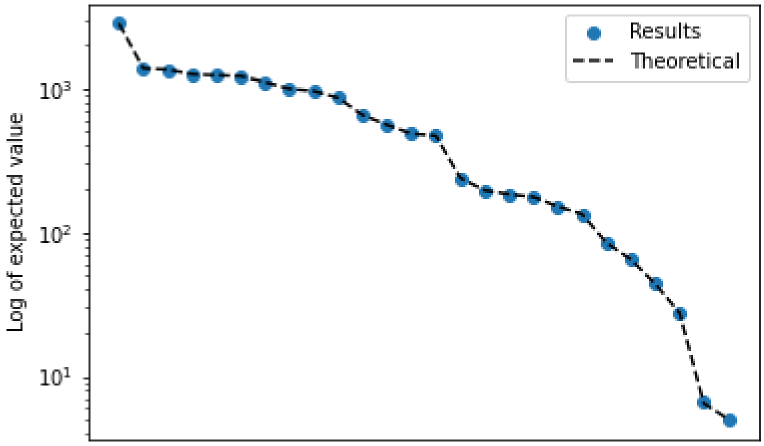

Fig. 10: Comparison between the mean value of each letter counter and the theoretical expected value, for the French text sample. Points were ordered from highest to lowest mean.

APPENDIX B - RELATIVE FREQUENCIES MEASURED BY THE APPROXIMATE COUNTERS WITH FIXED PROBABILITY

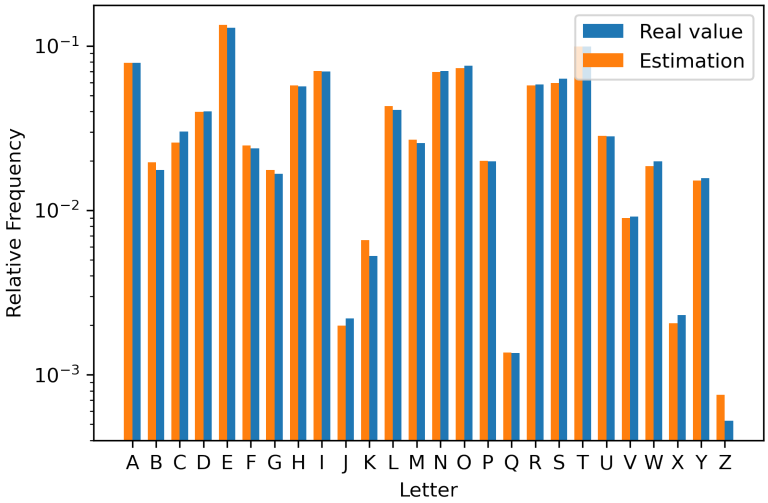

Fig. 11: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the counter, for the English text. The results are shown in logarithmic scale to better visualize lower frequencies.

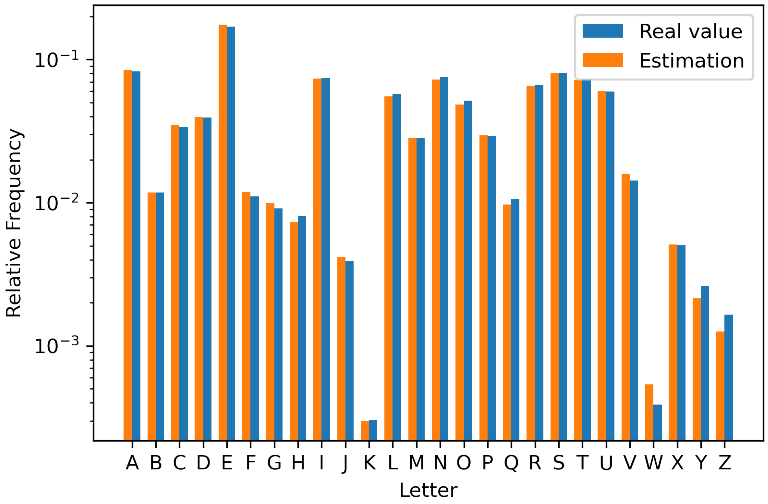

Fig. 12: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the counter, for the French text. The results are shown in logarithmic scale to better visualize lower frequencies.

APPENDIX C - TABLES FOR THE APPROXIMATE COUNTERS WITH FIXED PROBABILITY

TABLE III: Results of the approximate counter with fixed probability, ran 1000 times in the Portuguese text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
A	37584	37569.84	34960	39712	757.10	0.04	6.98
E	36029	36029.97	33296	38416	754.47	0.00	7.59
O	28623	28642.32	26592	31520	643.99	0.07	10.12
S	22100	22098.14	19680	24096	585.67	0.01	10.95
R	18555	18551.57	16768	20656	538.20	0.02	11.32
I	16980	16998.88	15296	18480	498.06	0.11	9.92
D	14444	14442.05	12896	15984	476.93	0.01	10.72
N	14216	14235.28	13024	15584	474.71	0.14	9.62
T	13684	13670.51	12368	15312	439.60	0.10	11.90
M	12923	12938.54	11584	14320	444.56	0.12	10.81
C	11270	11286.38	10192	12752	405.23	0.15	13.15
U	11252	11246.18	10032	12704	410.16	0.05	12.90
L	9073	9078.64	7792	10448	381.75	0.06	15.15
P	8339	8334.43	7056	9344	351.07	0.05	15.39
V	4301	4287.12	3600	5152	249.90	0.32	19.79
H	3474	3474.83	2832	4176	223.60	0.02	20.21
G	3250	3254.62	2656	4112	228.34	0.14	26.52
B	3183	3165.04	2496	3824	217.38	0.56	21.58
Q	3137	3130.37	2528	3824	214.92	0.21	21.90
F	2666	2674.66	2096	3360	190.14	0.32	26.03
Z	1107	1106.27	752	1504	126.35	0.07	35.86
J	849	851.34	528	1216	112.09	0.28	43.23
X	813	803.98	496	1168	113.60	1.11	43.67
Y	214	213.42	64	416	54.90	0.27	94.39
W	100	99.42	0	240	38.68	0.58	140.00
K	73	74.94	0	176	34.64	2.66	141.10

TABLE IV: Results of the approximate counter with fixed probability, ran 1000 times in the English text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
E	30350	30339.01	28048	32384	669.76	0.04	7.58
T	23325	23311.42	21696	25200	576.38	0.06	8.04
A	18575	18576.13	16704	20288	519.79	0.01	10.07
O	17888	17899.57	16304	19552	512.68	0.06	9.30
N	16591	16562.70	14624	17984	491.60	0.17	11.86
I	16470	16482.40	14736	18160	476.55	0.08	10.53
S	14875	14872.98	13648	16352	457.60	0.01	9.93
R	13723	13718.74	11920	15552	447.04	0.03	13.33
H	13382	13396.72	12064	14640	451.82	0.11	9.85
L	9637	9644.26	8592	10992	376.62	0.08	14.06
D	9394	9383.14	8208	10384	369.01	0.12	12.63
C	7102	7101.74	6016	8128	329.14	0.00	15.29
U	6648	6642.19	5648	7872	324.54	0.09	18.41
M	6051	6056.83	4880	7136	298.47	0.10	19.35
F	5612	5623.79	4416	6480	283.01	0.21	21.31
W	4685	4681.46	3744	5488	266.40	0.08	20.09
P	4672	4659.79	3888	5520	261.32	0.26	18.15
B	4138	4148.32	3392	4896	244.19	0.25	18.32
G	3925	3929.12	3088	4656	235.57	0.10	21.32
Y	3693	3678.80	3056	4480	232.02	0.38	21.31
V	2159	2159.10	1632	2704	173.61	0.00	25.24
K	1243	1238.27	848	1680	140.62	0.38	35.16
X	544	541.01	256	848	88.25	0.55	55.88
J	520	516.80	240	816	85.59	0.62	56.92
Q	319	317.54	80	528	68.93	0.46	74.92
Z	124	125.60	16	288	43.92	1.29	132.26

TABLE V: Results of the approximate counter with fixed probability, ran 1000 times in the French text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
E	45127	45115.65	42528	47760	817.52	0.03	5.83
A	21969	21972.58	20208	23920	565.41	0.02	8.88
S	21579	21535.97	19584	23216	582.64	0.20	9.25
N	20058	20057.31	18352	21680	543.45	0.00	8.51
I	19715	19711.73	17824	21520	551.99	0.02	9.59
T	19480	19476.46	17760	21232	540.10	0.02	8.99
R	17698	17677.82	16128	19136	495.11	0.11	8.87
U	15944	15943.04	14128	17280	502.29	0.01	11.39
L	15359	15359.18	13744	16768	492.04	0.00	10.52
O	13799	13810.86	12176	15456	464.36	0.09	12.01
D	10500	10501.38	9344	11840	396.21	0.01	12.76
C	8964	8952.99	7856	10064	358.40	0.12	12.36
P	7784	7800.58	6944	8880	344.58	0.21	14.08
M	7507	7506.88	6336	8560	326.65	0.00	15.60
V	3820	3808.56	3152	4480	236.05	0.30	17.49
B	3133	3117.58	2544	3808	210.50	0.49	21.54
F	2947	2948.90	2304	3712	209.57	0.06	25.96
Q	2818	2823.22	2240	3440	202.95	0.19	22.07
G	2423	2424.22	1792	3088	183.54	0.05	27.45
H	2142	2138.37	1552	2672	181.10	0.17	27.54
X	1351	1345.38	912	1888	145.55	0.42	39.75
J	1039	1041.81	592	1488	131.77	0.27	43.21
Y	702	708.42	400	1088	101.60	0.91	54.99
Z	441	439.26	224	704	79.89	0.39	59.64
W	104	105.95	0	240	38.62	1.88	130.77
K	81	80.10	0	208	35.07	1.12	156.79

APPENDIX D - EXPECTED VALUE OF THE MORRIS COUNTERS

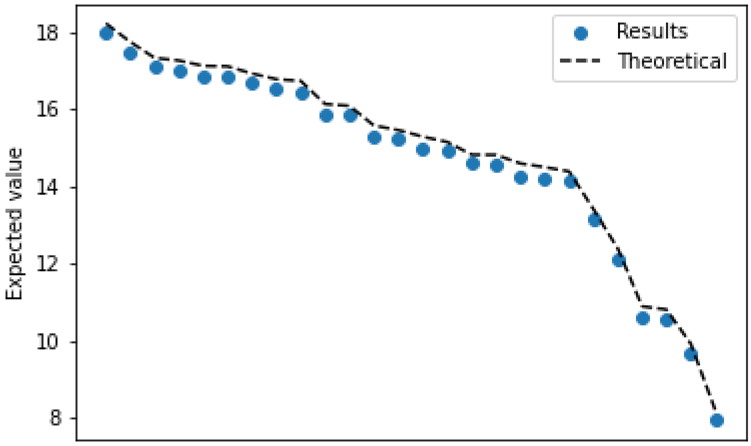

Fig. 13: Comparison between the mean value of each letter counter and the theoretical expected value, for the English text sample. Points were ordered from highest to lowest mean.

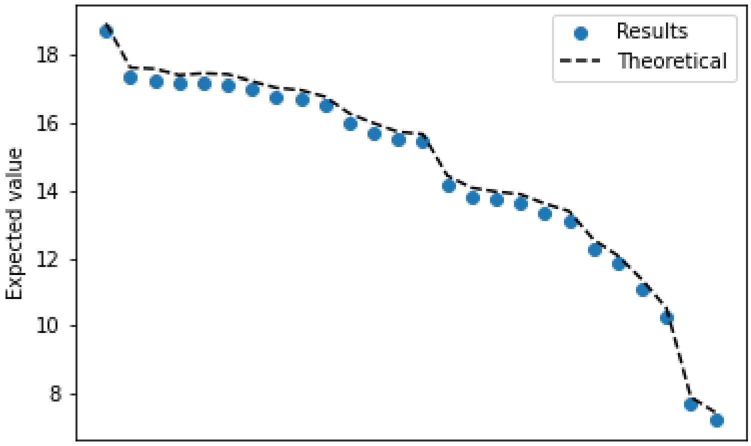

Fig. 14: Comparison between the mean value of each letter counter and the theoretical expected value, for the French text sample. Points were ordered from highest to lowest mean.

APPENDIX E - RELATIVE FREQUENCIES MEASURED BY THE MORRIS COUNTERS

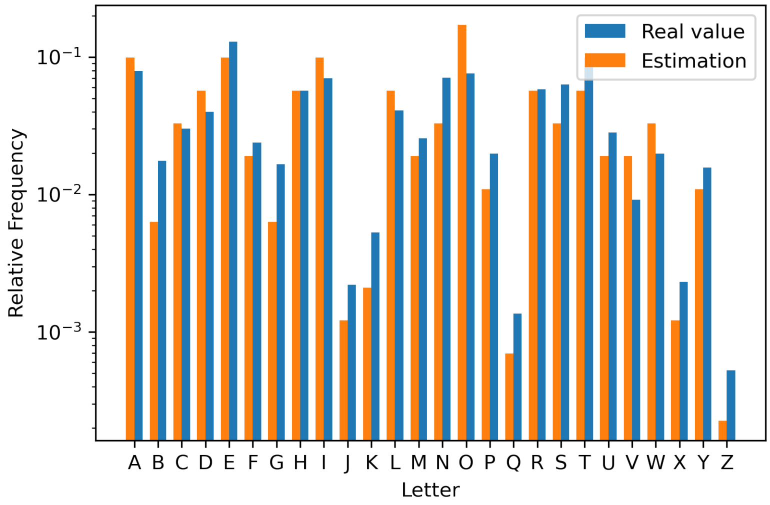

Fig. 15: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the counter, for the English text. The results are shown in logarithmic scale to better visualize lower frequencies.

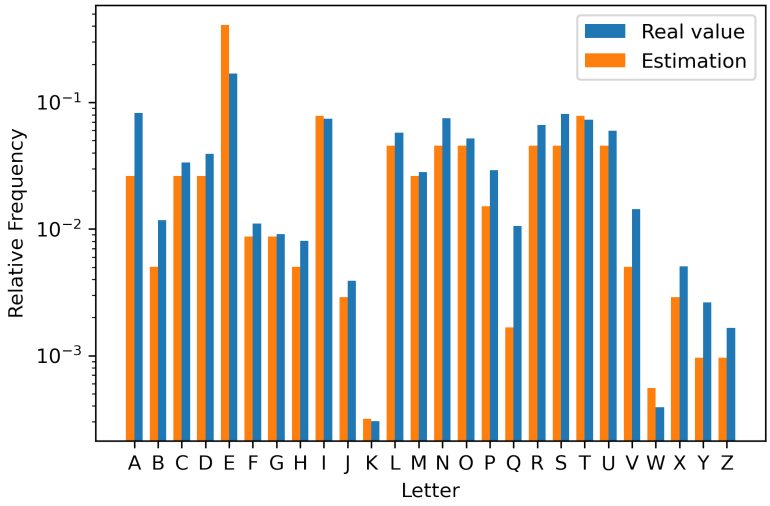

Fig. 16: Comparison between the real relative frequency of each letter counter and the value estimated in a random run of the counter, for the French text. The results are shown in logarithmic scale to better visualize lower frequencies.

APPENDIX F - TABLES FOR THE MORRIS COUNTERS

TABLE VI: Results of the Morris counter, ran 1000 times in the Portuguese text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
A	37584	32782.28	8960.13	241984.94	24905.30	12.78	543.85
E	36029	31424.52	8960.13	241984.94	22122.90	12.78	571.64
O	28623	23720.01	5172.13	139709.07	15860.34	17.13	388.10
S	22100	19420.90	5172.13	80660.07	12420.39	12.12	264.98
R	18555	15892.16	2985.13	80660.07	10420.62	14.35	334.71
I	16980	14340.53	2985.13	80660.07	9915.00	15.54	375.03
D	14444	12645.18	2985.13	80660.07	8310.34	12.45	458.43
N	14216	12377.12	1722.47	80660.07	8909.70	12.94	467.39
T	13684	12028.51	2985.13	80660.07	8534.40	12.10	489.45
M	12923	11587.41	2985.13	80660.07	8661.17	10.33	524.16
U	11252	9799.62	2985.13	80660.07	7116.20	12.91	616.85
C	11270	9681.88	2985.13	46568.11	6299.28	14.09	313.20
L	9073	7724.75	1722.47	46568.11	5668.91	14.86	413.26
P	8339	7016.54	1722.47	46568.11	4618.75	15.86	458.44
V	4301	3632.38	993.47	15521.13	2447.82	15.55	260.87
H	3474	3026.44	572.58	15521.13	2055.85	12.88	346.78
G	3250	2944.39	572.58	26885.11	2158.10	9.40	727.23
B	3183	2723.27	572.58	15521.13	2045.84	14.44	387.63
Q	3137	2687.57	572.58	15521.13	1978.42	14.33	394.78
F	2666	2328.29	572.58	15521.13	1687.16	12.67	482.19
Z	1107	971.82	189.28	5172.13	680.26	12.21	367.22
J	849	731.78	189.28	5172.13	497.26	13.81	509.20
X	813	687.22	108.28	2985.13	456.66	15.47	267.17
Y	214	180.94	34.52	993.47	120.56	15.45	364.24
W	100	84.76	18.93	993.47	64.89	15.24	893.47
K	73	62.94	9.93	329.58	45.48	13.79	351.48

TABLE VII: Results of the Morris counter, ran 1000 times in the English text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
E	30350	26840.84	5172.13	139709.07	20030.71	11.56	360.33
T	23325	20060.70	5172.13	139709.07	13753.67	13.99	498.97
A	18575	16289.95	2985.13	80660.07	11267.01	12.30	334.24
O	17888	15470.05	5172.13	80660.07	10200.00	13.52	350.92
N	16591	14348.41	2985.13	139709.07	11698.24	13.52	742.08
I	16470	14324.78	2985.13	80660.07	9667.03	13.03	389.74
S	14875	13119.33	2985.13	80660.07	9006.34	11.80	442.25
R	13723	11962.61	2985.13	139709.07	9302.02	12.83	918.07
H	13382	11448.20	2985.13	80660.07	8002.94	14.45	502.75
L	9637	8393.01	1722.47	46568.11	5732.18	12.91	383.22
D	9394	8342.43	1722.47	46568.11	6256.34	11.19	395.72
C	7102	5999.41	1722.47	46568.11	4327.39	15.53	555.70
U	6648	5862.53	1722.47	46568.11	3767.64	11.82	600.48
M	6051	5157.94	1722.47	26885.11	3410.26	14.76	344.31
F	5612	4914.43	993.47	26885.11	3414.85	12.43	379.06
W	4685	4128.66	993.47	26885.11	2973.90	11.87	473.86
P	4672	4050.00	993.47	26885.11	2964.52	13.31	475.45
B	4138	3451.36	572.58	26885.11	2465.76	16.59	549.71
G	3925	3302.87	993.47	15521.13	2322.47	15.85	295.44
Y	3693	3209.80	993.47	26885.11	2393.28	13.08	628.00
V	2159	1888.18	572.58	8960.13	1385.09	12.54	315.01
K	1243	1064.25	329.58	5172.13	707.58	14.38	316.10
X	544	466.06	108.28	1722.47	307.09	14.33	216.63
J	520	451.12	108.28	2985.13	340.57	13.25	474.06
Q	319	275.15	61.52	1722.47	193.35	13.74	439.96
Z	124	106.23	18.93	993.47	80.54	14.33	701.18

TABLE VIII: Results of the Morris counter, ran 1000 times in the French text.

Letter	Exact counts	Avg estim.	Min estim.	Max estim.	Std. dev.	Avg rel. error (%)	Max. rel. error (%)
E	45127	39644.83	8960.13	241984.94	27774.60	12.15	436.23
A	21969	18842.86	5172.13	139709.07	13492.60	14.23	535.94
S	21579	17845.40	5172.13	139709.07	12096.02	17.30	547.43
T	19480	17096.85	2985.13	80660.07	11418.27	12.23	314.07
N	20058	17087.46	5172.13	80660.07	11961.65	14.81	302.13
I	19715	16569.76	2985.13	139709.07	11674.78	15.95	608.64
R	17698	15487.06	2985.13	80660.07	10776.80	12.49	355.76
U	15944	13581.37	2985.13	80660.07	8910.71	14.82	405.90
L	15359	13315.39	2985.13	80660.07	9746.85	13.31	425.16
O	13799	11864.42	2985.13	80660.07	8465.67	14.02	484.54
D	10500	9039.24	1722.47	46568.11	6013.15	13.91	343.51
C	8964	7615.18	1722.47	46568.11	5186.33	15.05	419.50
P	7784	6890.45	1722.47	46568.11	4519.65	11.48	498.25
M	7507	6677.98	1722.47	46568.11	5016.48	11.04	520.33
V	3820	3321.08	572.58	26885.11	2359.63	13.06	603.80
B	3133	2705.36	572.58	15521.13	2014.02	13.65	395.41
F	2947	2581.84	329.58	8960.13	1656.70	12.39	204.04
Q	2818	2447.75	572.58	15521.13	1693.46	13.14	450.79
G	2423	2112.35	572.58	8960.13	1498.46	12.82	269.79
H	2142	1805.90	329.58	8960.13	1251.81	15.69	318.31
X	1351	1169.93	329.58	8960.13	816.66	13.40	563.22
J	1039	907.18	189.28	5172.13	634.41	12.69	397.80
Y	702	608.73	108.28	2985.13	417.53	13.29	325.23
Z	441	386.26	61.52	1722.47	259.11	12.41	290.58
W	104	92.46	18.93	572.58	66.46	11.10	450.56
K	81	70.24	18.93	329.58	48.40	13.28	306.89