Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 425 a 442

On the Entropy of Written Spanish

Sobre la entropía del español escrito

Fabio G. Guerrero[a]

Escuela de Ingeniería Eléctrica y Electrónica, Facultad de Ingeniería, Universidad del Valle, Cali, Colombia

Abstract

A discussion on the entropy of the Spanish language by means of a practical method for calculating the entropy of a text by direct computer processing is presented. As an example of application, thirty samples of Spanish text are analyzed, totaling 22.8 million characters. Symbol lengths from $n = 1$ to 500 were considered for both words and characters. Both direct computer processing and the probability law of large numbers were employed for calculating the probability distribution of the symbols. An empirical relation on entropy involving the length of the text (in characters) and the number of different words in the text is presented. Statistical properties of the Spanish language when viewed as produced by a stochastic source, (such as origin shift invariance, ergodicity and asymptotic equipartition property) are also analyzed.

***Key words*:** Law of large numbers, Shannon entropy, Stochastic process, Zipf's law.

Resumen

Se presenta una discusión sobre la entropía de la lengua española por medio de un método práctico para el cálculo de la entropía de un texto mediante procesamiento informático directo. Como un ejemplo de aplicación, se analizan treinta muestras de texto español, sumando un total de 22,8 millones de caracteres. Longitudes de símbolos desde $n = 1$ hasta 500 fueron consideradas tanto para palabras como caracteres. Para el cálculo de la distribución de probabilidad de los símbolos se emplearon procesamiento directo por computador y la ley de probabilidad de los grandes números. Se presenta una relación empírica de la entropía con la longitud del texto (en caracteres) y el número de palabras diferentes en el texto. Se analizan también propiedades estadísticas de la lengua española cuando se considera como producida por una fuente estocástica, tales como la invarianza al desplazamiento del origen, ergodicidad y la propiedad de equipartición asintótica.

***Palabras clave*:** entropía de Shannon, ley de grandes números, ley de Zipf, procesos estocásticos.

[a]Assistant Professor. E-mail: fabio.guerrero@correounivalle.edu.co

1. Introduction

Spanish is a language which is used by more than four hundred million people in more than twenty countries, and it has been making its presence increasingly felt on the Internet (Marchesi 2007). Yet this language has not been as extensively researched at entropy level. The very few calculations which have been reported have been obtained, as for most languages, by indirect methods, due in part to the complexity of the problem. Having accurate entropy calculations for the Spanish language can thus be considered a pending task. Knowing the value of H, in general for any language, is useful for source coding, cryptography, language space dimension analysis, plagiarism detection, and so on. Entropy calculation is at the lowest level of language analysis because it only takes into account source symbol statistics and their statistical dependence, without any further consideration of more intelligent aspects of language such as grammar, semantics, punctuation marks (which can considerably change the meaning of a sentence), word clustering, and so on.

Several approaches have been devised for several decades for finding the entropy of a language. Shannon (1948) initially showed that one possible way to calculate the entropy of a language, H, would be through the limit $H = \lim_{n\to\infty} -\frac{1}{n}H(B_i)$, where B_i is a sequence of n symbols. Finding H using methods such as the one suggested by this approach is difficult since it assumes that the probability of the sequences, $p(Bi)$, is an asymptotically increasing function of n, as n tends to infinity. Another difficulty posed by this approach is that an extremely large sample of text would be required, one that considered all possible uses of the language. Another suggested way to calculate H is by taking $H = \lim_{n\to\infty} F_n$, where $F_n = H(j|B_i) = H(B_ij) - H(B_i)$. B_i is a block of n-1 symbols, j is the symbol next to B_i, $H(j|B_i)$ is the conditional entropy of symbol j given block B_i. In this approach, the series of approximations $F_1, F_2, \ldots$ provides progressive values of conditional entropy. F_n, in bits/symbol, measures the amount of information in a symbol considering the previous $n-1$ consecutive symbols, due to the statistics of the language. The difficulty of using these previous methods in practice was put under evidence when in his pioneering work Shannon (1951) used instead a human prediction approach for estimating the entropy of English, getting 0.6 and 1.3 bits/letter as bounds for printed English, considering 100-letter sequences. Gambling estimations have also been used, providing an entropy estimation of 1.25 bits per character for English (Cover & King 1978). The entropy rate of a language could also be estimated using ideal source coders since, by definition, this kind of coder should compress to the entropy limit. A value of 1.46 bits per character has been reported for the entropy of English by means of data compression (Teahan & Cleary 1996). The entropy of the fruit fly genetic code has been estimated using universal data compression algorithms (Wyner, Jacob & Wyner 1998). As for the Spanish language, values of 4.70, 4.015, and 1.97 bits/letter for F_0, F_1, and F_W respectively were reported (Barnard III 1955) using an extrapolation technique on frequency data obtained from a sample of 6,513 different words. F_W is the entropy, in bits/letter, based on single-word frequency.

Another venue for finding H has been based on a purely mathematical framework derived from stochastic theory, such as the one proposed by Crutchfield & Feldman (2003). Unfortunately, as the same authors recognize it, it has lead, in practice, to very limited results for finding the entropy of a language. In general, as all these results suggest, finding the entropy of a language by classic methods has proved to be a challenging task. Despite some remarkable findings in the past decades, the search for a unified mathematical model continues to be an open problem (Debowski 2011).

In the past it was implicitly believed that attempting to find the average uncertainty content of a language by direct analysis of sufficiently long samples could be a very difficult task to accomplish. Fortunately, computer processing capacity available at present has made feasible tackling some computing intensive problems such as the search in large geometric spaces employed in this work. Michel, Shen, Aiden, Veres, Gray, Team, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak & Aiden (2011) discuss, as an example of this trend, the use of huge computational resources to research the relationship between linguistics and cultural phenomena. This paper is organized as follows: In Section 2 the methodology used to obtain all the values reported is discussed; in Section 3 the results of the observations are presented; Section 4 presents a discussion and analysis of the most relevant results and, finally, in Section 5 the main conclusions of this work are summarized. All the samples and support material used in this work are publicly available at http://sistel-uv.univalle.edu.co/EWS.html. Aspects such as the analysis of grammar, semantics, and compression theory are beyond the scope of this paper.

2. Methodology

Thirty samples of literature available in Spanish were chosen for this study. Tables 1 and 2 show the details of the samples and its basic statistics. The works used in this paper as samples of written Spanish were obtained from public libraries available on the Internet such as librodot [1] and the virtual library Miguel de Cervantes [2]. The selection of the samples was done without any particular consideration of publication period, author's country of origin, and suchlike. A file of news provided to the author by the Spanish press agency EFE was also included in the samples for analysis. The selected material was processed using a T3500 Dell workstation with 4 GB RAM. The software used to do the all the calculations presented in this work was written in Mathematica® 8.0. For simplicity, a slight preprocessing was done on each sample, leaving only printable characters. Strings of several spaces were reduced to one character and the line feed control character (carry return) was replaced by a space character, allowing for fairer comparisons between samples. The samples were character encoded using the ISO 8859-1 standard (8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1) which has 191 characters from the Latin script, providing a full set of charac-

[1] http://www.librodot.com
[2] http://www.cervantesvirtual.com

ters for the Spanish language. For instance, the $\tilde{n}$ letter corresponds to 0xf1, etc. The total amount of characters of the thirty samples in table 1 is 22,882,449 and the total amount of words is 4,024,911. The rounded average for the number of different one-character symbols (uppercase, lowercase, and punctuation marks) for the thirty samples was 93. The reason we consider the distinction between uppercase and lowercase symbols is that when characterizing an information source at entropy level, lowercase and uppercase symbols produce different message vectors from the transmission point of view (e.g. the word *HELLO* produces a completely different message vector than the word *hello*).

Table 1: Set of Text Samples

Sample	Name	Author
1	La Biblia	Several authors
2	efe-B2	EFE Press agency
3	Amalia	José Mármol
4	Crimen y Castigo	Fyodor Dostoevsky
5	Rayuela	Julio Cortázar
6	Doña Urraca de Castilla	F. Navarro Villoslada
7	El Corán	Prophet Muhammad
8	Cien Años de Soledad	Gabriel García Márquez
9	La Araucana	Alonso de Ercilla
10	El Papa Verde	Miguel Angel Asturias
11	América	Franz Kafka
12	La Altísima	Felipe Trigo
13	Al Primer Vuelo	José María de Pereda
14	Harry Potter y la Cámara Secreta	J.K. Rowling
15	María	Jorge Isaacs
16	Adiós a las Armas	Ernest Hemingway
17	Colmillo Blanco	Jack London
18	El Alférez Real	Eustaquio Palacios
19	Cañas y Barro	Vicente Blasco Ibáñez
20	Aurora Roja	Pío Baroja
21	El Comendador Mendoza	Juan C. Valera
22	El Archipiélago en Llamas	Jules Verne
23	Doña Luz	Juan Valera
24	El Cisne de Vilamorta	Emilia Pardo Bazán
25	Cuarto Menguante	Enrique Cerdán Tato
26	Las Cerezas del Cementerio	Gabriel Miró
27	Tristana	Benito Pérez Galdós
28	Historia de la Vida del Buscón	Francisco de Quevedo
29	El Caudillo	Armando José del Valle
30	Creció Espesa la Yerba	Carmen Conde

In Table 2 the parameter α is the average word length, given by $\sum L_i p_i$, where L_i and p_i are the length in characters and the probability of the i-th word respectively. The weighted average of α for the whole set of samples is 4.491 letters per word. The word dispersion ratio, WDR, is the percentage of different words over the total number of words.

The values of entropy were calculated using the entropy formula $\sum p_i \log_2 p_i$. The frequency of the different symbols (n-character or n-word symbols) and the law

of large numbers were used to find the symbol probabilities as $p_i \approx n_i/n_{total}$. First, we started considering word symbols, since words are the constituent elements of the language. However, a more refined analysis based on characters was also carried out. Entropy values for both n-character and n-word symbols from n=1 to 500 were calculated. Considering symbols up to a length of five hundred was a suitable number for practical proposes, this will be discussed in the next section.

Table 2: Sample Details

Sample	Number of Characters	Alphabet Size (A_S)	Number of Words	Different Words	WDR(%)	α
1	5722041	100	1049511	40806	3.89	4.27
2	1669584	110	279917	27780	9.92	4.80
3	1327689	88	231860	18871	8.14	4.51
4	1215215	91	207444	17687	8.53	4.63
5	984129	117	172754	22412	12.97	4.50
6	939952	84	161828	17487	10.81	4.58
7	884841	93	160583	12236	7.62	4.32
8	805614	84	137783	15970	11.59	4.73
9	751698	82	129888	15128	11.65	4.63
10	676121	93	118343	16731	14.14	4.45
11	594392	88	101904	11219	11.01	4.66
12	573399	89	98577	14645	14.86	4.53
13	563060	82	100797	13163	13.06	4.35
14	528706	89	91384	10884	11.91	4.60
15	499131	87	88376	12680	14.35	4.45
16	471391	91	81803	10069	12.31	4.49
17	465032	91	81223	10027	12.35	4.58
18	462326	89	82552	10699	12.96	4.43
19	436444	79	75008	10741	14.32	4.66
20	393920	90	68729	10598	15.42	4.47
21	387617	86	69549	10289	14.79	4.38
22	363171	88	61384	8472	13.80	4.73
23	331921	83	59486	9779	16.44	4.41
24	312174	77	53035	11857	22.36	4.65
25	304837	87	49835	12945	25.98	4.95
26	302100	75	51544	10210	19.81	4.64
27	299951	82	52571	10580	20.13	4.48
28	232236	74	42956	7660	17.83	4.23
29	224382	83	36474	7470	20.48	5.00
30	159375	81	27813	6087	21.89	4.48

One worthwhile question at this point is "does entropy change when changing the origin point in the sample?". For this purpose, we calculated entropy values considering symbols for different shifts from the origin for non overlapping symbols, as illustrated by figure 1, for the case of trigrams.

It can easily be seen that, for symbols of length n, symbols start repeating (i.e., symbols are the same as for shift=0, except for the first one) after n shifts. As a result, the number of individual entropy calculations when analyzing symbols from length $n = 1$ up to k was $\frac{k(k+1)}{2}$. For the $k = 500$ case used in this work, this gives 125,250 individual entropy calculations for every sample analyzed. The

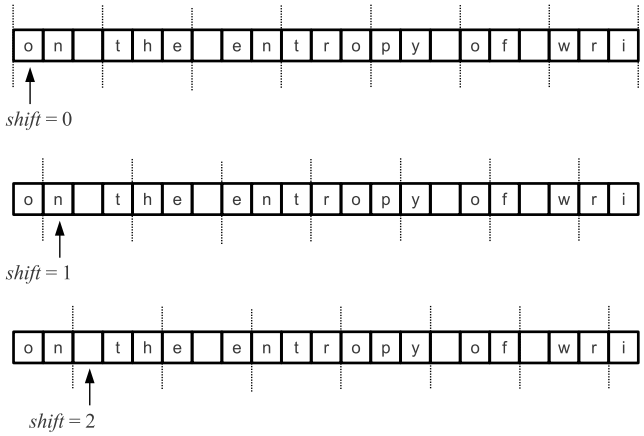


FIGURE 1: Origin invariance analysis.

individual shift entropies so obtained were then averaged for every n. Values of n for which the maximum value of entropy was produced were identified, as well as values of n from which all symbols present in the text become equiprobable with reasonable certainty, i.e., none of them repeat more than once in the sample.

3. Results

3.1. Entropy Values Considering Words

Figure 2 shows the values of average entropy for n-word symbols. For ease of display, only data for samples 1, 2, 12 and 30 and $n = 1$ to 20 are shown. The rest of the literary works exhibited the same curve shapes with values in between. All the analyzed samples exhibited invariance to origin shift. For example, for sample 8 (*Cien Años de Soledad*) the values for $n = 4$ were: 15.024492 ($shift = 0$), 15.028578 ($shift = 1$), 15.025693 ($shift = 2$), 15.027212 ($shift = 3$). This means that $P(w_1, ..w_L) = P(w_{1+s}, ..w_{L+s})$ for any integer s, where $\{w_1, ..w_L\}$ is a L-word sequence. This is a very useful property to quickly find the entropy of a text it because it makes necessary to compute values for just one shift thus reducing the process to a few seconds for practical purposes.

Also since the weighted value for 1-word entropy for the set analyzed was 10.0064 bits/character, the weighted value of F_W is therefore 2.23 bits/character.

3.2. Entropy Values Considering n-Character Symbols

Figure 3 shows the averaged entropy values for n-character symbols. Again for ease of display, only data for samples 1, 2, 12 and 30 and $n = 1$ to 100 are shown. All samples also exhibited the origin shift invariance property. For example, for sample 8 (*Cien Años de Soledad*), the values of entropy for $n = 4$ characters were:

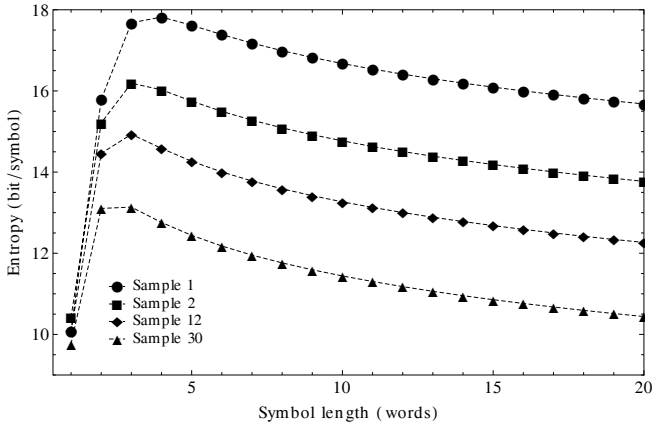


FIGURE 2: Entropy for n-word symbols for samples 1, 2, 12 and 30.

12.267881 ($shift = 0$), 12.264343 ($shift = 1$), 12.268751 ($shift = 2$), 12.269691 ($shift = 3$). Therefore, $P(c_1, .., c_L) = P(c_{1+s}, .., c_{L+s})$ for any integer s. As in the case of words, the rest of literary works exhibited the same curve shapes with values in between.

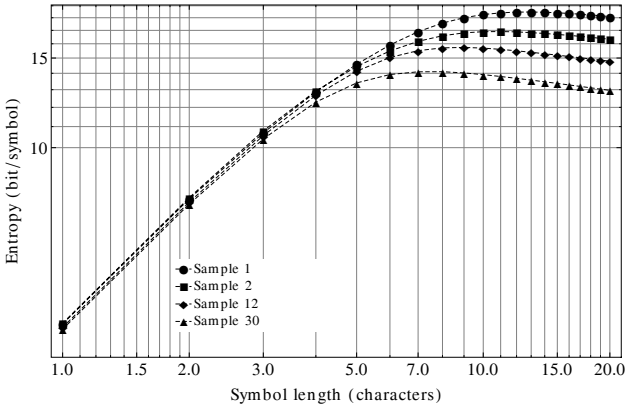


FIGURE 3: Log Plot of entropy for n-character symbols for samples 1, 2, 12 and 30.

3.3. Processing Time

Figure 4 shows the processing time of every sample for both words and characters for *all* shifts of n ($1 \leq n \leq 500$), that is, 125,250 entropy calculations for each sample. Due to the origin shift invariance property, only calculations for one shift (for instance $shift = 0$) are strictly required thus reducing the time substantially. For example, the processing time of sample 1 for only one shift was 433 seconds while the processing time for sample 30 was just nine seconds. Analysis for all shifts of n were done in this work in order to see if entropy varied when changing

the point of origin in the text. A carefully designed algorithm based on Mathematica's sorting functions was employed to obtain the probability of symbols, however, a discussion on the optimality of this processing is beyond the scope of this paper.

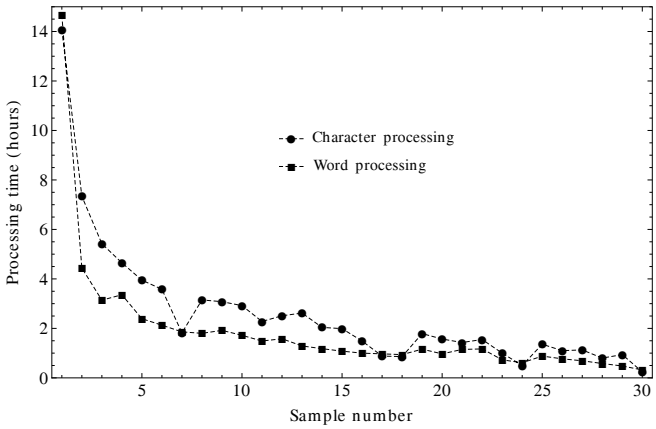


FIGURE 4: Processing time considering *all* shifts of n (125,250 entropy calculations)

3.4. Reverse Entropy

If we take the text in reverse order, for instance "yportne eht no" instead of "on the entropy", it is possible to evaluate the reverse conditional entropy, that is, the effect of knowing how much information can be gained about a previous character when later characters are known. It was observed that entropy of the reverse text carried out for the same set of samples produced exactly the same values as for the forward entropy case. This was first observed by Shannon for the case of the English language in his classical work (Shannon 1951) on English prediction.

4. Discussion

4.1. Frequency of Symbols and Entropy

Figure 5 shows a plot of the fundamental measure function of information, $p_i \log_2 p_i$, which is at the core of the entropy formula. This function has its maximum, 0.530738, at $p_i = 0.36788$. Therefore, infrequent symbols, as well as very frequent symbols, add very little to the total entropy. This should not be confused with the value of $p_i = \frac{1}{n}$ that produces the maximum amount of entropy for a probability space with n possible outcomes. The entropy model certainly has some limitations because entropy calculation is based solely on probability distribution. In fact, two different texts with very different location of words can have the same entropy, yet one of them can lead to a very much more efficient source encoding than the other.

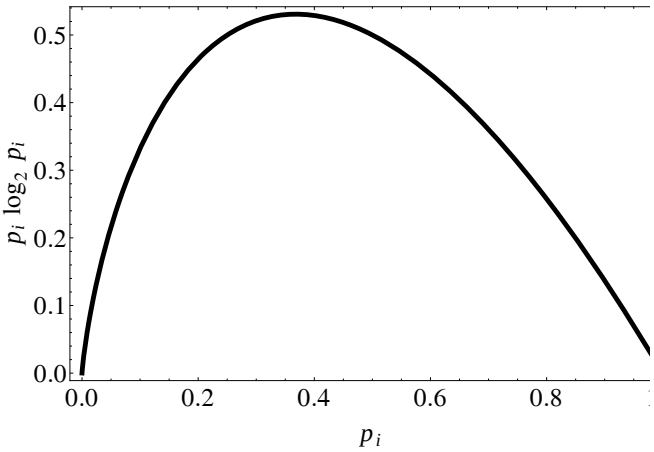


FIGURE 5: Fundamental function of information.

4.2. Log-log Plots and the Spanish Language Constant

The fact that the basic statistical properties of entropy are essentially the same for short length symbols regardless of the sample (and the entropy is similar for any shift of the origin) means it is possible to use a sufficiently long sample, for instance sample 2, to study the Spanish language constant. Figure 6 shows the log-log plot for sample 2 which contained 82,656 different 3-word symbols, 79,704 different 2-word symbols, and 27,780 different 1-word symbols. Log-log plots for the rest of samples were found to be similar to those of figure 6, at least for 2-word and 1-word symbols.

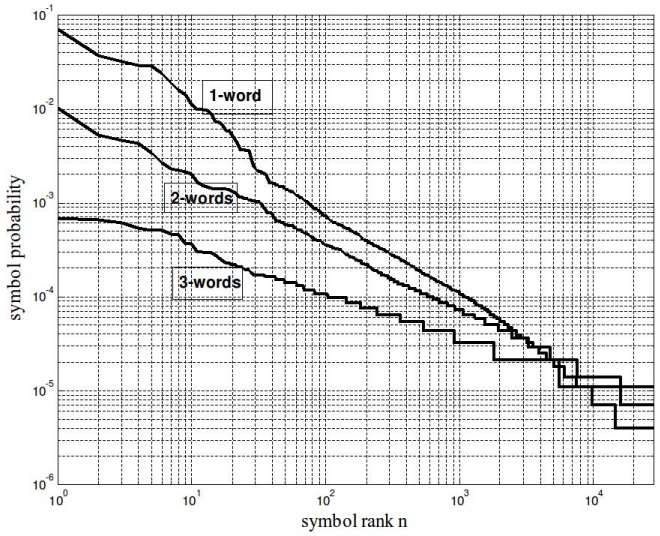


FIGURE 6: Symbol rank versus n-word probability in Sample 2.

Smoothing the 1-word curve in figure 6, the probability of the r-th most frequent 1-word symbol is close to $0.08/r$, assuming r is not too large. This behavior corresponds to the celebrated Zipf law first presented in 1939 (Zipf 1965) which nowadays some authors also call the Zipf-Mandelbrot law (Debowski 2011). Figure 7 shows the log-log plot for for sample 2 which contained 14,693 different trigrams, 2,512 different digrams, and 110 different characters; all values considered for $shift = 0$. Log-log plots for the rest of the samples were found to be similar to those of figure 7. Even when a distinction between upper case and lower case symbols is made in this work, no significant difference was found with the constant obtained when analyzing the database of the 81,323 most frequent words (which makes no distinction between upper case and lower case symbols). This database was compiled by Alameda & Cuetos (1995) from a corpus of 1,950,375 words of written Spanish.

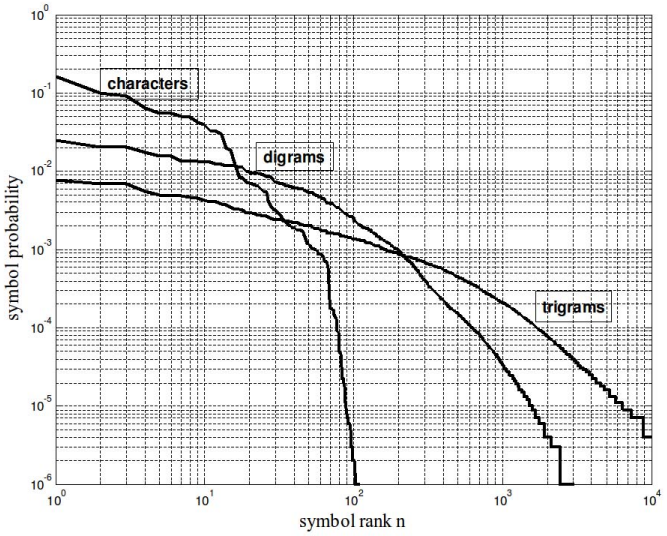


FIGURE 7: Symbol rank versus n-character probability for Sample 2.

4.3. Conditional Entropy

We now evaluate the uncertainty content of a character given some previous text. Initially F_0, in bits per character, is given by $\log_2(A_S)$, where A_S is the alphabet size. F_1 takes into account single-character frequencies and it is given by $F_1 = \sum_i p_i \log_2 p_i$. F_2 considers the uncertainty content of a character given the previous one:

$$F_2 = -\sum_{i,j} p(i,j) \log_2 p(j|i) = -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p_i \log_2 p_i \qquad (1)$$

Similarly, F_3 gives the entropy of a character given the previous two characters (digram):

$$F_3 = -\sum_{i,j,k} p(i,j,k)\log_2 p(k|ij) = -\sum_{i,j,k} p(i,j,k)\log_2 p(i,j,k) + \sum_{i,j} p_{i,j}\log_2 p_{i,j} \tag{2}$$

and so on. Table 3 shows, for simplicity, values for F_n from F_1 to F_{15} only, rounded to two significant digits.

TABLE 3: Conditional Entropy F_n

	n														
S_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4.51	3.43	2.76	2.18	1.72	1.33	0.98	0.68	0.43	0.26	0.14	0.06	0.01	-0.02	-0.04
2	4.52	3.46	2.82	2.13	1.52	1.05	0.69	0.42	0.22	0.09	0.01	-0.03	-0.06	-0.06	-0.07
3	4.39	3.34	2.73	2.11	1.57	1.11	0.72	0.42	0.21	0.08	-0.01	-0.05	-0.07	-0.08	-0.08
4	4.43	3.39	2.74	2.09	1.54	1.07	0.67	0.37	0.18	0.06	-0.01	-0.05	-0.06	-0.07	-0.08
5	4.40	3.41	2.81	2.16	1.55	1.02	0.59	0.30	0.12	0.01	-0.05	-0.08	-0.08	-0.09	-0.09
6	4.39	3.35	2.74	2.13	1.56	1.05	0.62	0.32	0.13	0.02	-0.04	-0.07	-0.08	-0.09	-0.09
7	4.46	3.31	2.57	1.93	1.40	0.96	0.61	0.36	0.20	0.09	0.03	-0.02	-0.04	-0.06	-0.06
8	4.27	3.27	2.67	2.06	1.50	1.02	0.63	0.34	0.16	0.05	-0.02	-0.06	-0.07	-0.08	-0.08
9	4.32	3.28	2.70	2.11	1.58	1.06	0.61	0.29	0.10	-0.01	-0.06	-0.09	-0.10	-0.10	-0.09
10	4.40	3.36	2.78	2.16	1.51	0.95	0.51	0.22	0.05	-0.04	-0.08	-0.09	-0.10	-0.10	-0.09
11	4.33	3.32	2.66	2.00	1.43	0.93	0.54	0.28	0.11	0.01	-0.04	-0.07	-0.08	-0.09	-0.09
12	4.44	3.38	2.74	2.11	1.46	0.90	0.47	0.20	0.03	-0.05	-0.09	-0.10	-0.10	-0.10	-0.10
13	4.36	3.31	2.71	2.07	1.47	0.93	0.52	0.23	0.07	-0.03	-0.07	-0.09	-0.09	-0.10	-0.09
14	4.44	3.40	2.69	1.98	1.35	0.84	0.47	0.22	0.08	-0.01	-0.05	-0.07	-0.08	-0.09	-0.09
15	4.38	3.33	2.72	2.07	1.43	0.87	0.46	0.20	0.05	-0.04	-0.08	-0.09	-0.10	-0.10	-0.09
16	4.46	3.35	2.69	2.00	1.37	0.83	0.44	0.19	0.05	-0.03	-0.07	-0.08	-0.09	-0.09	-0.09
17	4.32	3.30	2.63	1.98	1.39	0.89	0.50	0.24	0.07	-0.01	-0.06	-0.08	-0.09	-0.09	-0.09
18	4.35	3.33	2.71	2.05	1.41	0.86	0.45	0.19	0.04	-0.04	-0.08	-0.09	-0.10	-0.09	-0.09
19	4.29	3.29	2.64	1.98	1.37	0.87	0.49	0.23	0.08	-0.01	-0.06	-0.08	-0.09	-0.09	-0.09
20	4.44	3.37	2.73	2.03	1.34	0.78	0.37	0.14	0.01	-0.06	-0.08	-0.10	-0.10	-0.10	-0.09
21	4.37	3.33	2.71	2.04	1.37	0.79	0.39	0.13	0.00	-0.05	-0.09	-0.09	-0.11	-0.06	-0.12
22	4.38	3.34	2.65	1.91	1.26	0.75	0.40	0.17	0.05	-0.03	-0.05	-0.08	-0.08	-0.08	-0.08
23	4.34	3.30	2.67	2.00	1.35	0.78	0.38	0.13	0.01	-0.06	-0.09	-0.10	-0.10	-0.10	-0.09
24	4.38	3.36	2.78	2.08	1.34	0.71	0.30	0.06	-0.05	-0.09	-0.11	-0.11	-0.11	-0.10	-0.10
25	4.32	3.37	2.80	2.09	1.32	0.69	0.29	0.06	-0.05	-0.09	-0.10	-0.11	-0.11	-0.10	-0.10
26	4.42	3.35	2.71	2.01	1.28	0.71	0.32	0.10	-0.03	-0.07	-0.10	-0.11	-0.10	-0.10	-0.10
27	4.37	3.34	2.74	2.06	1.33	0.72	0.31	0.08	-0.04	-0.09	-0.11	-0.11	-0.11	-0.10	-0.10
28	4.33	3.25	2.63	1.94	1.26	0.71	0.32	0.10	-0.03	-0.09	-0.10	-0.11	-0.11	-0.10	-0.10
29	4.28	3.28	2.62	1.89	1.21	0.68	0.32	0.11	0.00	-0.07	-0.09	-0.10	-0.10	-0.09	-0.09
30	4.40	3.35	2.66	1.89	1.11	0.52	0.17	0.01	-0.08	-0.11	-0.12	-0.11	-0.11	-0.11	-0.10

We observe in table 3 that, at some point, conditional entropies become negative. Although $H(X,Y)$ should always be greater or equal to $H(Y)$, the estimation on conditional entropy in this study becomes negative because the length of the text is not sufficiently long, in contrast to the required condition of the theoretical model $n \to \infty$. This behavior has also been observed in the context of bioinformatics and linguistics (Kaltchenko & Laurier 2004). The following example should help to clarify the explanation. Let's consider first the following text in Spanish which has 1000 characters: ⟨⟨ *Yo, señora, soy de Segovia. Mi padre se llamó Clemente Pablo, natural del mismo pueblo; Dios le tenga en el cielo. Fue, tal como todos dicen, de oficio barbero, aunque eran tan altos sus pensamientos que se corría de que le llamasen así, diciendo que él era tundidor de mejillas y sastre de barbas. Dicen que era de muy buena cepa, y según él bebía es cosa para creer. Estuvo casado con Aldonza de San Pedro, hija de Diego de San Juan y nieta de Andrés de San Cristóbal. Sospechábase en el pueblo que no era cristiana vieja, aun viéndola con canas y rota, aunque ella, por los nombres y sobrenombres de sus pasados, quiso esforzar que era descendiente de la gloria. Tuvo muy buen parecer para letrado; mujer de amigas y cuadrilla, y de pocos enemigos, porque hasta los tres del alma no los tuvo por tales; persona de valor y*

conocida por quien era. Padeció grandes trabajos recién casada, y aun después, porque malas lenguas daban en decir que mi padre metía el dos de bastos para sacar el as de oros⟩⟩. This text has 250 four-character symbols (e.g. {Yo, },{seño},{ra, }) with 227 of them being different. Factorizing common probability terms we find: $H_{4-char} = 209(\frac{1}{250}\log_2\frac{1}{250})+14(\frac{1}{125}\log_2\frac{1}{125})+3(\frac{3}{250}\log_2\frac{3}{250})+\frac{2}{125}\log_2\frac{2}{125} = 7.76$ bits/symbol. This text has 200 five-character symbols (e.g.{Yo, s},{eñora},{, soy}) with 192 being different five-character symbols. Factorizing common probability terms we find: $H_{5-char} = 184(\frac{1}{200}\log_2\frac{1}{200}) + 8(\frac{1}{100}\log_2\frac{1}{100}) = 7.56$ bits/symbol. Thus the entropy of a character given the previous four characters are know and would be $H(X|Y) = H_{5-char} - H_{4-char} = -0.20$ bits/character. For sample 1 (which has 5,722,040 characters) a similar behavior is observed: The greatest number of different symbols (418,993) occurs for n=10 (572,204 total 10-character symbols) for which H=18.26 bits/symbol. The highest entropy, 18.47 bits/symbol, is produced by 13-character symbols (there are 440,156 total 13-character symbols, and 395,104 different 13-character symbols). For 14-character symbols (408,717 total; 378,750 different) the entropy is 18.45 bits/symbol. Then the entropy of a character given the previous thirteen characters are know, in this case, is $18.45 - 18.47 = -0.02$ bits/character. With increasing n, the probability distribution tends to become uniform and H starts decreasing monotonically with n, as shown in figure 3 of the paper. When the symbols in the sample become equiprobable the value of H is given by $\log_2\lfloor\frac{\text{total number of characters}}{n}\rfloor$. Again, these seemingly paradoxical values are explained by the differences between mathematical models and real world, as well as the assumptions on which they are based[3].

4.4. Entropy Rate and Redundancy

To estimate the entropy rate, a polynomial interpolation of third degree is first applied to the values of F_n. As an example, figure 8 shows the interpolated curves for samples one and thirty.

Figure 8 shows that F_n becomes negative after crossing by zero, and from this point asymptotically approaches zero as $n \to \infty$. Therefore,

$$\lim_{n\to\infty} F_n = \lim_{n\to N_Z} F_n \tag{3}$$

In equation 3, N_Z is the root of the interpolated function F_n. The n-character entropy values of figure 3 are also interpolated to find H_{NZ}, the entropy value corresponding to N_Z. The redundancy is given by $R = \frac{H_L}{H_{max}}$, where H_L is the source's entropy rate, and $H_{max} = \log_2(A_S)$. Finally, the value of H_L is calculated as $H_L \approx \frac{H_{NZ}}{N_Z}$. Table 4 summarizes the values of N_Z, H_{NZ}, H_L, and R. It should be clear that the previous interpolation process is used to get a finer approximation to the value of entropy. Just as in thermodynamics a system in equilibrium state produces maximum entropy, equation 3 captures the symbol distribution that produces the highest level of entropy (or amount of information) in the text.

[3]An insightful dissertation on real world and models is presented in Slepian (1976).

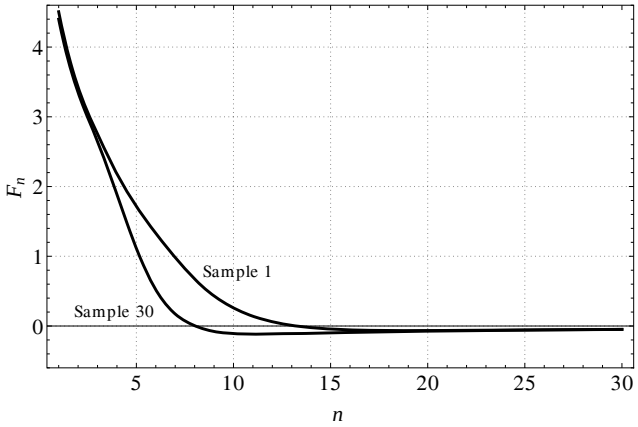


FIGURE 8: Interpolated curves of conditional entropy (bits/character) for samples 1 and 30.

In Table 4, the weighted average of H_L is 1.54 bits/character. Since the weighted average of the alphabet size in Table 1 is 92.98 characters the average redundancy, R, for the analyzed sample set, comprising nearly 23 million characters, is:

$$R = 1 - \frac{1.54}{\log_2 92.98} \approx 76.486\%$$

Taking $H(X)$ equal to 1.54 bits/character, for a text of Spanish of 140 characters, there would exist $2^{nH(X)} \approx 7.98 \times 10^{64}$ typical sequences. Because the roots of F_n occur at small values of n and, as it has been observed this method permits to find the value of entropy in a very short time (analysis for only one shift, for instance *shift*=0, is required). As it can be observed in Table 4, in general, a sample with lower WDR has more redundancy, the opposite also being true. In general, and as a consequence of Zipf's, law the greater the size of a sample, the smaller its WDR. An interesting empirical relation found in this work involving H_L, the length of the text (in characters) L, and the number of different words (V) in the text is:

$$H_L \approx \frac{2.192}{\log_V L} \tag{4}$$

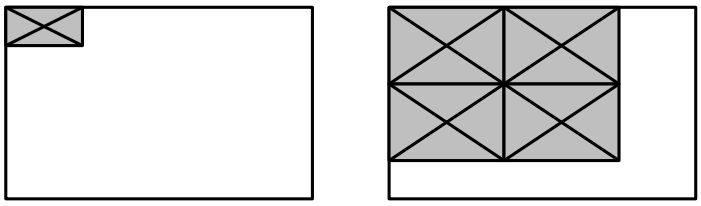
Equation 4 indicates that texts with small word dictionaries (compared to the length of the text in characters) have smaller H_L because there is higher redundancy. This corroborates the well known fact that larger documents are more compressible than smaller ones. The compression factor[4] using bzip compression for samples 1 and 30 is 0.25 and 0.33 respectively, which is in total agreement with sample 1 having more redundancy than sample 30. Equation 4 is a reasonable approximation considering that in this work L takes into consideration punctuation

[4]The compression factor is defined in this work as the size after compression over the size before compression.

Table 4: Entropy Rate and Redundancy

Sample	N_Z	H_{NZ}	H_L	$R(\%)$
1	13.23	18.47	1.40	78.99
2	11.23	16.91	1.51	77.80
3	10.92	16.67	1.53	76.38
4	10.82	16.53	1.53	76.54
5	10.10	16.36	1.62	76.43
6	10.22	16.28	1.59	75.08
7	11.54	15.91	1.38	78.92
8	10.60	15.95	1.50	76.46
9	9.90	16.05	1.62	74.49
10	9.48	15.93	1.68	74.31
11	10.14	15.61	1.54	76.16
12	9.31	15.73	1.69	73.92
13	9.64	15.64	1.62	74.47
14	9.92	15.46	1.56	75.94
15	9.49	15.50	1.63	74.64
16	9.55	15.36	1.61	75.28
17	9.79	15.30	1.56	75.98
18	9.44	15.37	1.63	74.86
19	9.81	15.23	1.55	75.36
20	9.11	15.19	1.67	74.32
21	9.01	15.14	1.68	73.85
22	9.60	14.90	1.55	75.98
23	9.06	14.96	1.65	74.12
24	8.47	14.99	1.77	71.77
25	8.48	14.94	1.76	72.66
26	8.72	14.89	1.71	72.58
27	8.58	14.93	1.74	72.61
28	8.71	14.53	1.67	73.14
29	9.01	14.40	1.60	74.93
30	8.05	14.10	1.75	72.38

marks. Figure 9 is intended to illustrate that as a sample has a higher WDR, there is a tendency to the equipartition of the sample space, increasing thus H_L.



a) b)

Figure 9: Illustration of Word Dispersion Ratio over word space: a) Lower WDR. b) Higher WDR.

The twenty-second version of the Dictionary of the Royal Academy of the Spanish Language (DRAS) has 88,431 lemmas (entries) with 161,962 definitions (i.e., meanings for the words according to the context in which they appear). If

compared to the total number of lemmas of the DRAS, the works analyzed in this work use a relatively small number of words. For instance, literature's Nobel Prize winner Gabriel García Márquez in his masterpiece, *Cien Años de Soledad*, used around sixteen thousand different words. Because the vocabulary at the end is finite, the WDR for larger texts has to be, in general, smaller.

Finally, when concatenating the whole set of thirty samples to form one larger sample (22.9 million characters) the results were: $\alpha = 4.491$ letter/word, $H_L = 1.496$ bits/character, and $R = 78.92\%$. The computing time ($shift = 0$) was thirty four minutes.

Many other samples of Spanish can be analyzed (for instance, science, sports, etc.) but Table 4 should give a good indication of what to expect in terms of the entropy for ordinary samples of written Spanish. However, as Table 4 also shows, finding an exact value for the entropy of Spanish is an elusive goal. We can only make estimations of entropy for particular text samples. The usefulness of the method presented here lies on its ability to provide a direct entropy estimation of a particular text sample.

4.5. Character Equiprobability Distance

We define the character equiprobability distance of a text sample, n_{aep}, as the value of n such that for any $n \geq n_{aep}$, all n-length symbols in the sample become equiprobable for all shifts of n. This means,

$$H = \log_2 \left\lfloor \frac{(\text{Total number of characters}) - shift}{n} \right\rfloor$$

for all $n \geq n_{aep}$. This definition demands symbol equiprobability for all shifts for every $n \geq n_{aep}$, in other words, every substring of length $n \geq n_{aep}$ only appears once, not matter its position in the text.

Table 5 shows the values of n_{aep} evaluated from $n = 1$ to 500 characters and $2^{n_{aep} H_L}$, the number of typical sequences of length n_{aep} characters. Plagiarism detection tools should take into account the value of n_{aep}, because for sequences shorter than n_{aep} characters, it is more likely to find similar substrings of text due to the natural restriction imposed by the statistical structure of the language. Large values of n_{aep} in Table 5 were found to be related to some text reuse such as, for instance, sample 2 where some partial news are repeated as part of a larger updated news report. As it is observed, the number of typical sequences is of considerable size despite the apparently small number of characters involved.

5. Conclusions

The evidence analyzed in this work shows that the joint probability distribution of Spanish does not change with position in time (origin shift invariance). Due to this property the method for finding the entropy of a sample of Spanish presented in this work is simple and computing time efficient. Both, a redundancy of 76.5%

Table 5: Equiprobable Distance ($1 \leq n \leq 500$)

Sample	n_{aep}	$2^{n_{aep}H_L}$
1	412	4.31E+173
2	452	2.88E+205
3	93	6.82E+042
4	53	2.57E+024
5	356	4.07E+173
6	124	2.24E+059
7	101	9.07E+041
8	76	2.08E+034
9	36	3.60E+017
10	116	4.62E+058
11	39	1.20E+018
12	255	5.36E+129
13	189	1.48E+092
14	84	2.80E+039
15	50	3.42E+024
16	61	3.67E+029
17	118	2.59E+055
18	208	1.15E+102
19	37	1.84E+017
20	43	4.14E+021
21	453	1.25E+229
22	69	1.57E+032
23	43	2.28E+021
24	29	2.83E+015
25	55	1.38E+029
26	38	3.64E+019
27	27	1.39E+014
28	32	1.22E+016
29	50	1.21E+024
30	43	4.49E+022

and a rate entropy of 1.54 bits/character were found for the sample set analyzed. A value of 2.23 bits/character was found for F_W. In general, lower values of WDR were observed for longer samples leading to higher values of redundancy, just in accordance with Zipf's law. Evidence also shows that, for every day texts of the Spanish language, $p(B_i)$ is not an asymptotically increasing function of n and the highest moment of uncertainty in a sample occurs for a relatively small value of n. Considering n-word symbols, $H_{\max}$ was found at a value of four or less words. When considering n-character symbols, $_{H\max}$ was found at a value of fourteen or less characters. An averaged value of n_{aep} close to 125 characters can be a good indication of how constrained we are by the statistical structure of the language. The probability of the r-th most frequent word in Spanish is approximately $0.08/r$. If compared to the constant of English, $0.1/r$, it can be concluded that the total probability of words in Spanish is spread among more words than in English. There is a clear indication of the relation between a text's dictionary size (number of different words) and H_L. In general, a text with a larger dictionary size causes H_L to increase. Texts with small word dictionaries

compared to the length of the text in characters have smaller H_L and thus should be more compressible. Since reverse entropy analysis produced exactly the same values as forward entropy, for prediction purposes the amount of uncertainty when predicting a text backwards is, despite being apparently more difficult, the same as predicting the text forwards. Finally, despite the fact that the basic statistical properties are similar regardless of the text sample analyzed, since entropy depends solely on probability distribution, every text of Spanish will exhibit its own value of entropy, thus making it difficult to talk about *the* entropy of Spanish.

Acknowledgment

The author would like to thank Ms. Ana Mengotti, edition board director of the EFE press agency in Bogota (Colombia), for the news archive provided for this research. Also thanks to the anonymous reviewers for their helpful comments.

[Recibido: noviembre de 2011 — Aceptado: septiembre de 2012]

References

Alameda, J. & Cuetos, F. (1995), 'Diccionario de las unidades lingüísticas del castellano, Volumen II: Orden por frecuencias'.
*http://www.uhu.es/jose.alameda

Barnard III, G. (1955), 'Statistical calculation of word entropies for four Western languages', *IEEE Transactions on Information Theory* **1**, 49–53.

Cover, T. & King, R. (1978), 'A convergent gambling estimate of the entropy of English', *IEEE Transactions on Information Theory* **IT-24**(6), 413–421.

Crutchfield, J. & Feldman, D. (2003), 'Regularities unseen, randomness observed: Levels of entropy convergence', *Chaos* **13**(6), 25–54.

Debowski, L. (2011), 'Excess entropy in natural language: Present state and perspectives', *Chaos* **21**(3).

Kaltchenko, A. & Laurier, W. (2004), 'Algorithms for estimating information distance with applications to bioinformatics and linguistics', *Canadian Conference Electrical and Computer Engineering* .

Marchesi, A. (2007), 'Spanish language, science and diplomacy (In Spanish)'. International Congress of the Spanish Language, Cartagena.
*http://corpus.canterbury.ac.nz

Michel, J., Shen, Y. K., Aiden, A., Veres, A., Gray, M., Team, T. G. B., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. & Aiden, E. (2011), 'Quantitative analysis of culture using millions of digitized books', *Science* **331**, 176–182.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423.

Shannon, C. E. (1951), 'Prediction and entropy of printed English', *Bell System Technical Journal* **30**, 47–51.

Slepian, D. (1976), 'On bandwidth', *Proceedings of the IEEE* **34**(3).

Teahan, W. & Cleary, J. (1996), 'The entropy of English using PPM-based models', *Data Compression Conference* pp. 53–62.

Wyner, A., Jacob, Z. & Wyner, A. (1998), 'On the role of pattern matching in information theory', *IEEE Transactions on Information Theory* **44**(6), 2045–2056.

Zipf, G. K. (1965), *The Psycho-Biology of Language: An Introduction to Dynamic Philology, Second Edition*, The MIT Press.