Supplementary Material

1 THE GRASSBERGER ENTROPY ESTIMATOR

Let us consider[1] the results based on the entropy estimator $\hat{H}_G$, proposed by Grassberger (2003), to those obtained from the "naive word entropy formula",

$$H(t) := \sum_{i=1}^{M} \hat{p}_i \log_2(\hat{p}_i), \tag{S1}$$

where for a text t, $\hat{p}_i := f_i/N$ is the relative frequency of a unique word w_i, f_i its total number of occurrences, N the total number of words in t, and M the size of the corresponding vocabulary, i.e. $N = \sum_{i=1}^{M} f_i$. Alternatively, $\hat{p}_i$ can be interpreted as the empirical probability to find w_i, sampled from an, a priori unknown, probability distribution. The normalised word entropy H_n, is then given by $H_n(t)/\log_2(M)$.

In Grassberger (2003), it is assumed that the following holds: the number of boxes (states) $M \gg 1$, the number $N \gg 1$ of particles distributed randomly and independently into the boxes, with n_i the random number of particles in box $i, i = 1, \ldots, M$, and $\sum_{i=1}^{M} n_i = N$.

Further, it is assumed that $M, N \to \infty$ and $n_i/N \to 0$ for all i; (which is not the case for small finite systems).

The final form of the entropy estimator $\hat{H}_G$ is obtained by combining formula (23) and equation (27) in Grassberger (2003):

$$\hat{H}_G = \ln(N) - \frac{1}{N} \sum_{i=1}^{M} n_i \left(\psi(n_i) + \frac{(-1)^{n_i}}{n_i(n_i+1)} \right), \tag{S2}$$

where $\psi(x) := \frac{d \ln(\Gamma(x))}{dx}$, is the digamma function.

In our case, M is equal to the size of the vocabulary underlying t, N to the number of words (tokens) in t, and n_i to the number of occurrences of vocabulary item i.

The normalised Grassberger entropy estimator $\hat{H}_{nG}$, is then given by $\hat{H}_G/\ln(M)$. Let us point out that for small samples $\hat{H}_{nG}$ does not satisfy the fundamental relation $\hat{H}_G \leq \ln(M)$.

1.1 Comparison of $\hat{H}_G$ vs H

In order to address the question on how much the outcome depends on the formula utilised, we performed the following additional analyses, using Python and the data visualisation library Seaborn, Waskom et al. (2017).

First, we calculated the normalised vocabulary entropy values, based on formulae (S1) and (S2), for every text in each corpus and then determined Spearman's rank correlation coefficient ρ, and Pearson correlation coefficients r, between the two resulting vectors. The results show that, according to both measures, the two quantities have a correlation coefficient almost equal to 1, cf. Table S3.

[1] We would like to thank the anonymous referee who pointed this out.

Corpus	a=slope	b=intercept
CA acts (EN)	1.0	-0.0
CA reg. (EN)	1.0	-0.0
CA acts (FR)	1.0	0.0
CA reg. (FR)	1.0	-0.0
F Codes (FR)	**1.1**	**-0.1**
D acts (DE)	**1.1**	-0.0
D reg. (DE)	1.0	0.0
UK PGA (EN)	1.0	-0.0
USC 1-54 (2020) (EN)	1.0	-0.0
U.S. CFR (2000) (EN)	**1.1**	-0.0
U.S. CFR (2019) (EN)	**1.1**	**-0.1**
CH Fed. acts (EN)	**1.1**	**-0.1**
CH Fed. reg. (EN)	1.0	-0.0
Shakespeare (EN)	1.0	-0.0
EP (DE)	**1.1**	**-0.1**
EP (EN)	**1.1**	-0.0
EP (FR)	1.0	-0.0

Table S1. Summary statistics showing the values of the slope a, and intercept b, of the linear regression equation $y = ax + b$, between the normalised vocabulary and normalised Grassberger entropy for each corpus; cutoff at 150K.

Then, using Seaborn, we determined the coefficients a (slope) and b (intercept) of the linear regression equation $y = ax + b$, between the normalised vocabulary and normalised Grassberger entropy, cf. Table S1.

In seven out of seventeen cases, the slope was equal to 1.1, and in all other ten cases equal to 1. The intercept was either ± 0.0, or, in four cases where $a = 1.1$, equal to -0.1. Therefore the Grassberger estimator should not change the average relative positions of the data points in the vertical direction, which is indeed the case, cf. Table S2.

Additionally, we plotted the normalised entropy vs the normalised Grassberger entropy for several data sets, cf. Figures S1, S2 and S3, which confirms the strong linear relation between the two quantities.

Third, by using Seaborn's kernel density estimation function, we determined the densities of the distributions of the normalised plain vocabulary and normalised Grassberger entropy, in order to visualise them, cf. Figure S4. As one observes, the respective graphs are related by a horizontal translation.

Finally, we re-plotted some of our comparisons in the complexity-entropy plane, which up to an overall vertical shift, genuinely reproduce the previous results, cf. Figure S5 and S6.

In conclusion, applying the Grassberger entropy estimator (S2) instead of the "plain" word entropy formula (S1), did not change the outcome of the present study, on the contrary, it strongly confirms it.

REFERENCES

Grassberger P. Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138* (2003).

[Dataset] Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0.8.1 (september 2017) (2017). doi:10.5281/zenodo.883859.

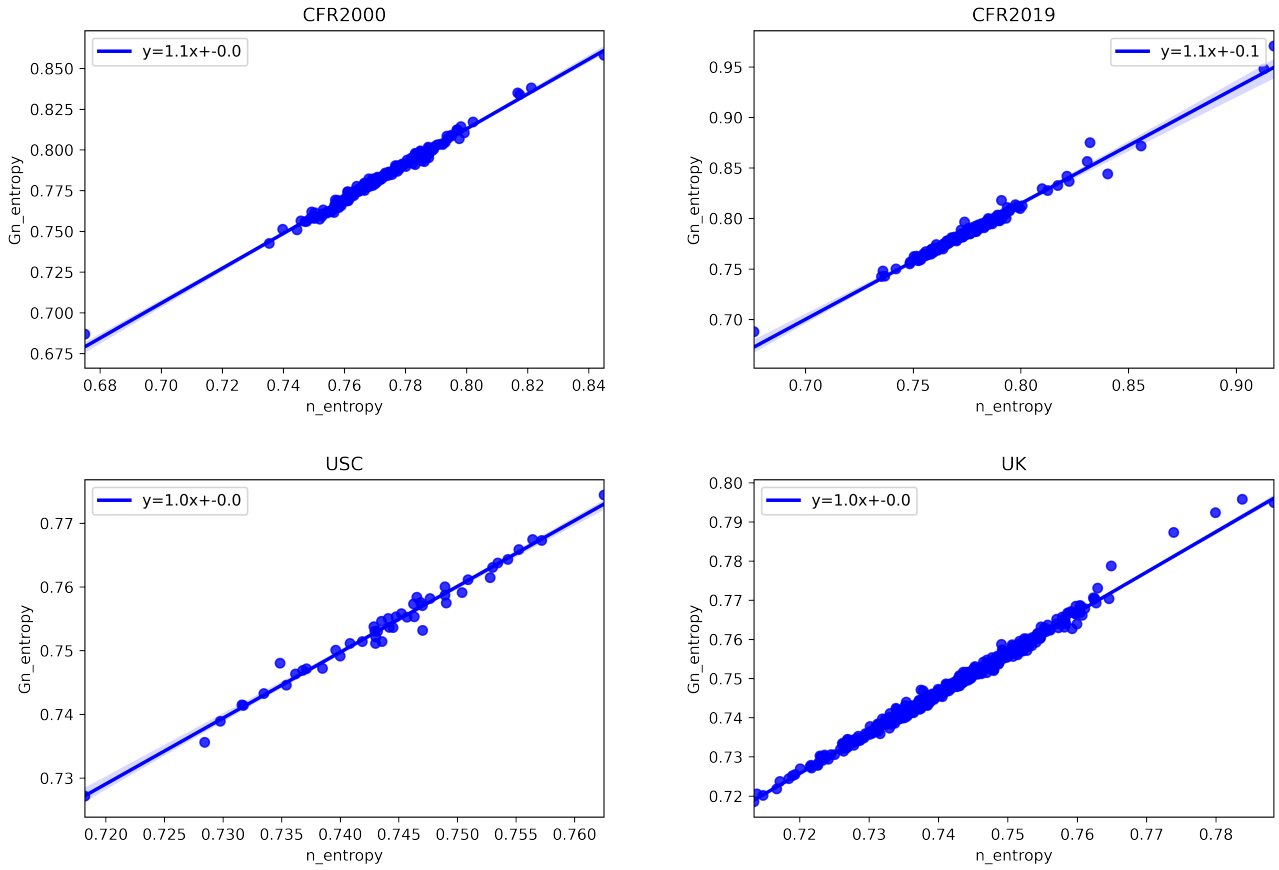

Figure S1. Figure showing the normalised vocabulary entropy vs the normalised Grassberger entropy and the linear regression, for selected corpora (CFR=Code of Federal Regulations). Top left: U.S. CFR 2000. Top right: U.S. CFR 2019. Bottom left: U.S. Code, Titles 1-54. Bottom right: UK Public General Acts (UKPGA). Cutoff at 150k.

Corpus	mean	std
CA acts (EN)	0.73	0.01
CA reg. (EN)	0.74	0.02
CA acts (FR)	0.76	0.01
CA reg. (FR)	0.75	0.02
F Codes (FR)	0.77	0.00
D acts (DE)	0.80	0.01
D reg. (DE)	0.81	0.01
UK PGA (EN)	0.75	0.01
USC 1-54 (2020) (EN)	0.75	0.00
U.S. CFR (2000) (EN)	0.78	0.02
U.S. CFR (2019) (EN)	0.79	0.02
CH Fed. acts (EN)	0.77	0.00
CH Fed. reg. (EN)	0.78	0.01
Shakespeare (EN)	0.81	0.00
EP (DE)	0.83	0.00
EP (EN)	0.78	0.00
EP (FR)	0.79	0.00

Table S2. Summary statistics showing the mean and standard deviation for the normalised Grassberger vocabulary entropy for all corpora. All texts are cutoff at 150K.

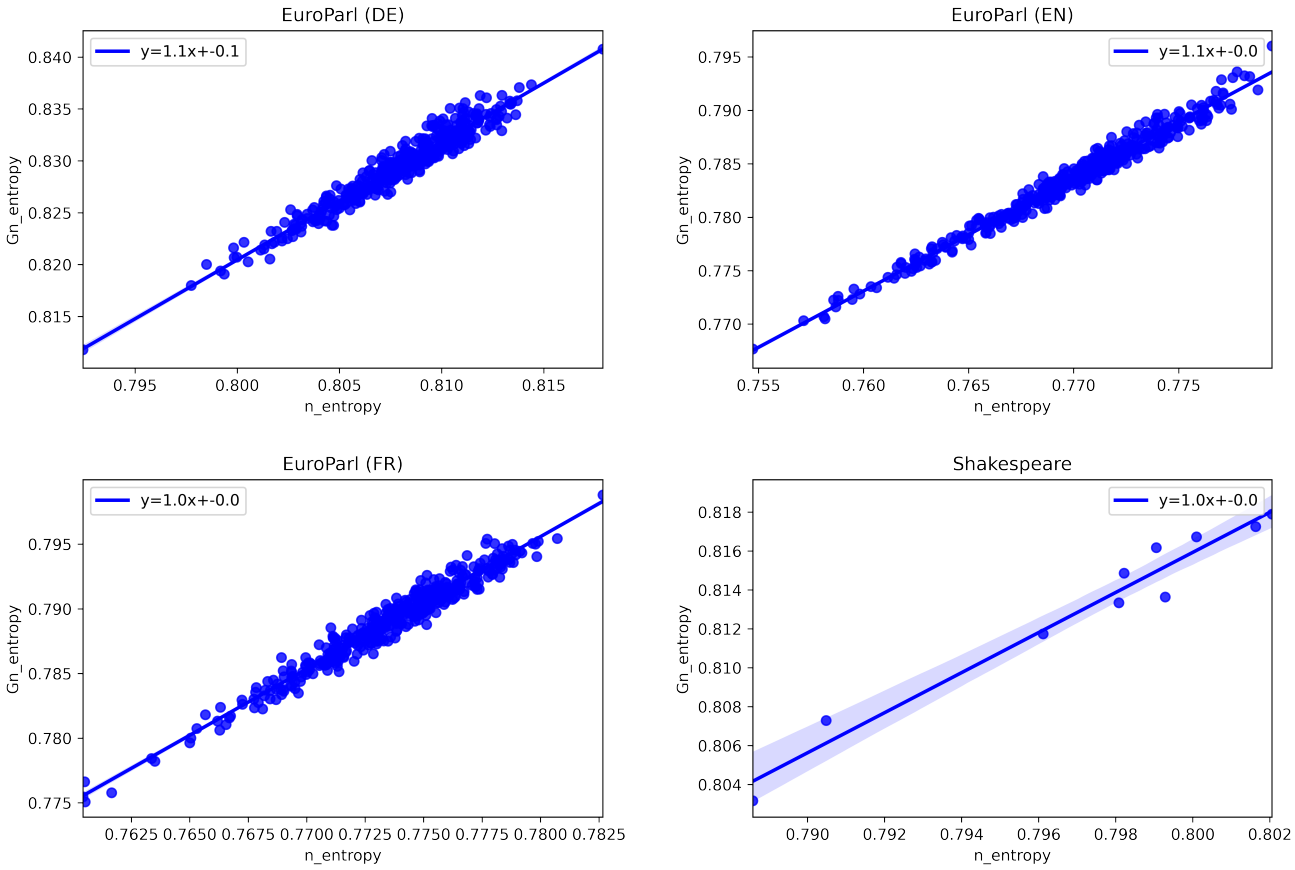

Figure S2. Figure showing the normalised vocabulary entropy vs the normalised Grassberger entropy and the linear regression, for the aligned translations of the German EuroParl corpus into English and French, and Shakespeare's collected works. Top left: EuroParl (German). Top right: EuroParl (English). Bottom left: EuroParl (French). Bottom right: Shakespeare (English). Cutoff at 150k.

Corpus	Pearson's r	Spearman's ρ
CA acts (EN)	0.990	0.976
CA reg. (EN)	0.997	0.995
CA acts (FR)	0.987	0.983
CA reg. (FR)	0.996	0.975
F Codes (FR)	0.987	0.970
D acts (DE)	0.971	0.983
D reg. (DE)	0.963	0.950
UK PGA (EN)	0.995	0.995
USC 1-54 (2020) (EN)	0.990	0.970
U.S. CFR (2000) (EN)	0.995	0.994
U.S. CFR (2019) (EN)	0.991	0.991
CH Fed. acts (EN)	0.994	1.000
CH Fed. reg. (EN)	0.995	1.000
Shakespeare (EN)	0.981	0.952

Table S3. Summary statistics showing the values of the Pearson correlation and the Spearman rank correlation coefficient between the normalised "raw" and Grassberger entropy estimates for each corpus; cutoff at 150K.

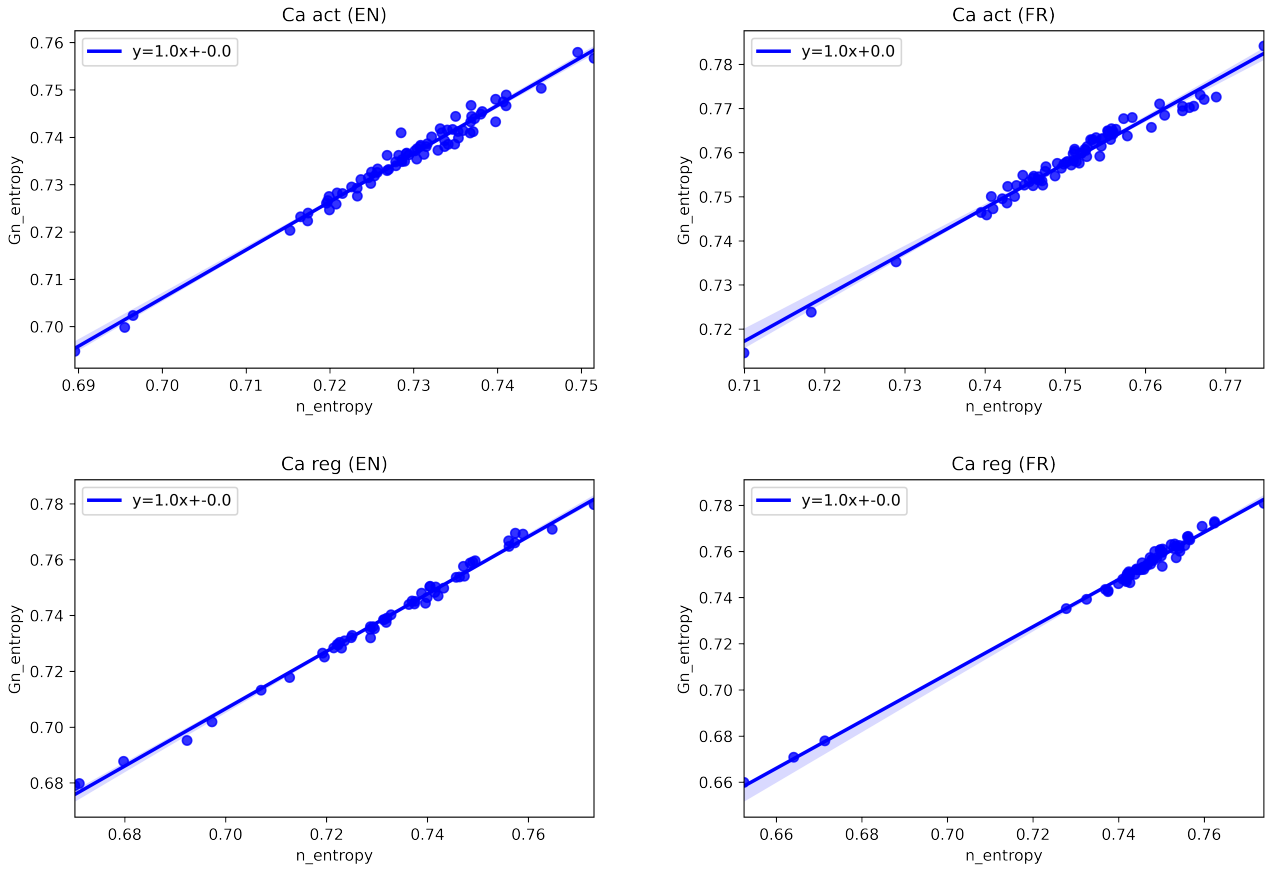

Figure S3. Figure showing the normalised vocabulary entropy vs the normalised Grassberger entropy and the linear regression, for the Canadian acts and regulations in English and French. Top left: Canadian acts (English). Top right: Canadian acts (French). Bottom left: Canadian regulations (English). Bottom right: Canadian regulations (French). Cutoff at 150k.

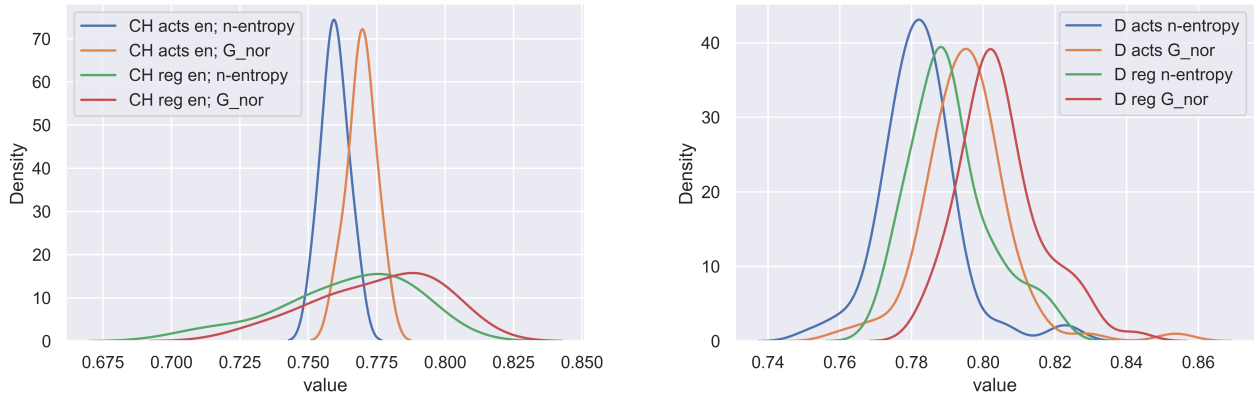

Figure S4. Kernel density estimates (KDE) of the plain normalised vocabulary and Grassberger entropies of the Swiss Federal acts and regulations in English (left), and the German acts and regulations in German (right).

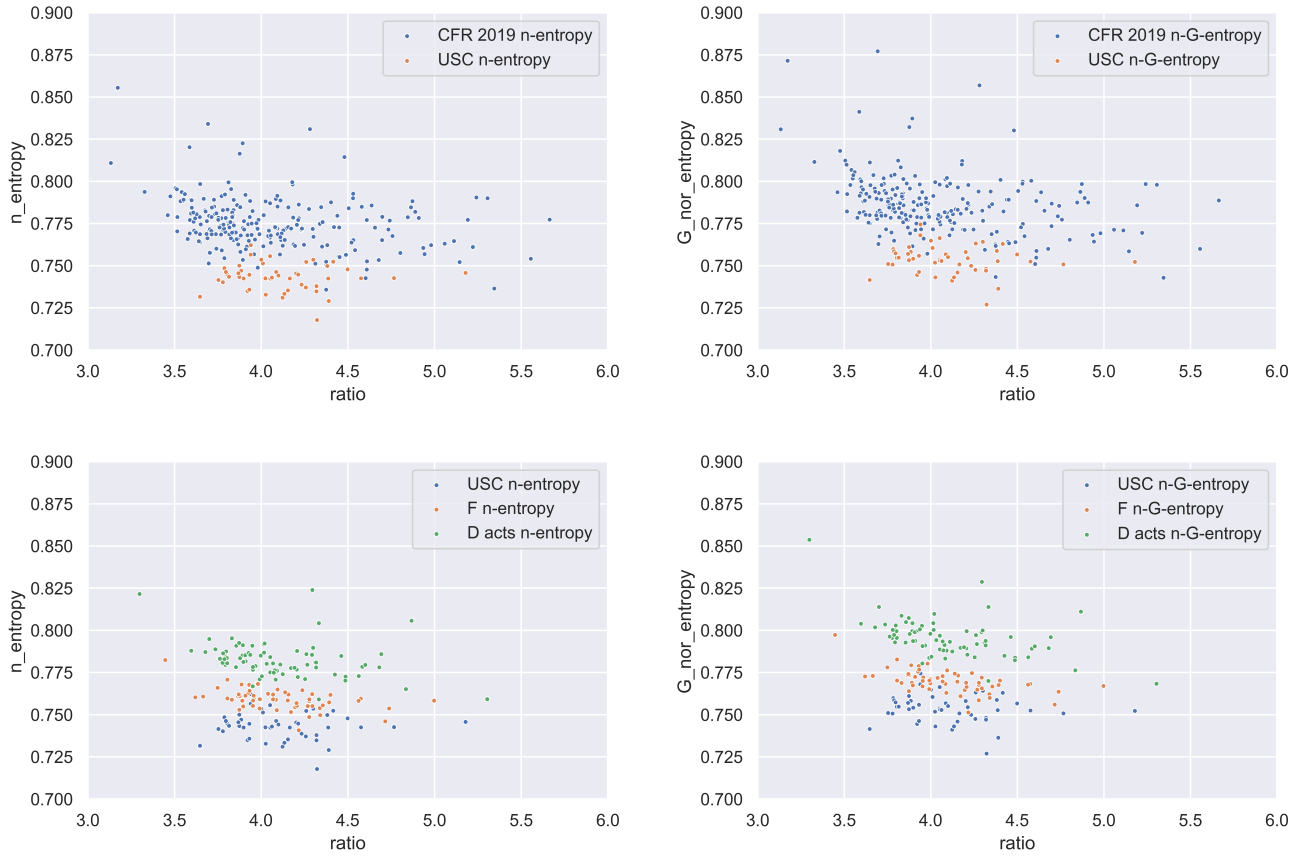

Figure S5. Plot of the compression factor (ratio) vs the normalised entropy / normalised Grassberger entropy estimator. CFR: Code of Federal Regulations. Top left: CFR 2019 and U.S. Code, Titles 1-54 for the normalised entropy. Top right: CFR 2019 and U.S. Code, Titles 1-54 for the normalised Grassberger entropy estimator. Bottom left: French acts (in French), German acts (in German) and U.S. Code, Titles 1-54 (in English) for the normalised entropy. Bottom right: French acts (in French), German acts (in German) and US Titles 1-54 (in English) for the normalised Grassberger entropy estimator. Cutoff at 150K.

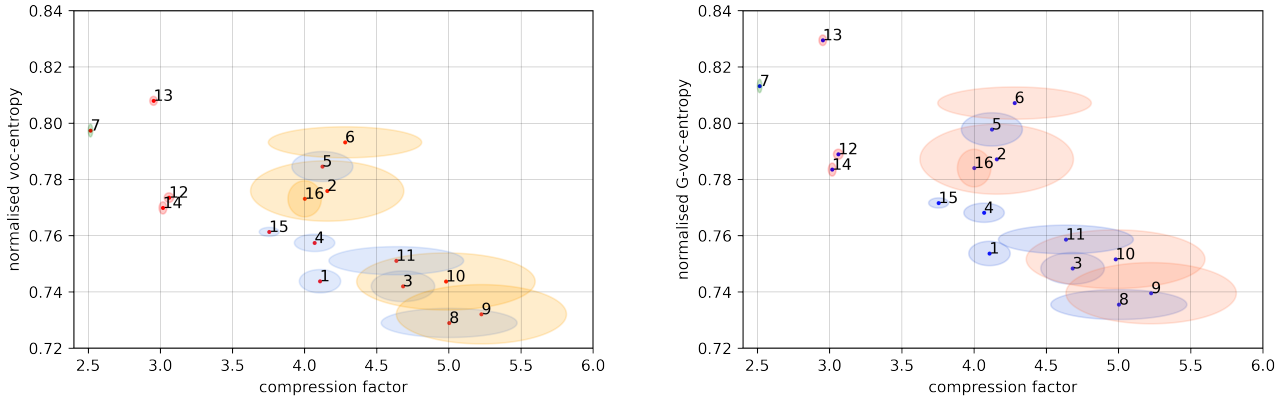

Figure S6. Figure showing the mean compression factor and mean normalised vocabulary entropy (left) and normalised Grassberger vocabulary entropy (right) for: 1=U.S. Code Titles 1-54, 2=U.S. CFR 2019, 3=UK general public acts, 4=French acts (FR), 5=German acts (DE), 6=German regulations (DE), 7=Shakespeare's collected works, 8=Canadian acts (EN), 9=Canadian regulations (EN), 10=Canadian regulations (FR), 11=Canadian acts (FR), 12=EuroParl speeches (FR), 13=EuroParl speeches (DE), 14=EuroParl speeches (EN), 15=Swiss Federal acts (EN), 16=Swiss Federal regulations (EN). The ellipses are centred around the mean values and have half-axes corresponding to $\sigma/2$ of the standard deviation of the compression factor and the normalised (Grassberger) vocabulary entropy, respectively. Colour code: left figure: red=speeches (EuroParl), green=literature (Shakespeare), light blue=acts, orange=regulations; right figure: red=speeches (EuroParl), green=literature (Shakespeare), dark light blue=acts, coral red=regulations; all texts truncated at 150K.