Originally published in: OSF Preprints, Center for Open Science (2024), 9pp.
DOI: *https://doi.org/10.31219/osf.io/p5nhd*

Corpus size strongly matters when analysing word frequency distributions

Alexander Koplenig[1]

1 Leibniz Institute for the German Language (IDS), Mannheim, Germany.

Correspondence: koplenig@ids-mannheim.de

In a previous study in this journal[1], Aceves and Evans present a large-scale quantitative information-theoretic analysis of parallel corpus data in ~1,000 languages to show that there are apparently strong associations between the way languages encode information into words and patterns of communication, e.g. the configuration of semantic information. During the peer review process, one reviewer raised the question of the extent to which the presented results depend on different corpus sizes (see the Peer Review File). This is a very important question given that most, if not all, of the quantities associated with word frequency distributions vary systematically with corpus size.[2–4] While Aceves and Evans claim that corpus size does not affect the results presented, I challenge this view by presenting reanalyses of the data that clearly suggest that it does.

To test for a potential bias due to corpus size, my reanalysis focusses on the relation between information density and semantic density (cf. Fig. 2 in the paper). Aceves and Evans interpret the obtained correlation to suggest a strong positive association between language information density and semantic density across corpora. I first concentrate on the "Bible_NT" corpus, which encompasses the majority of the languages investigated by Aceves and Evans, with 76.18% (number of languages $N_L = 761$) of the total 999 languages being exclusively available in this dataset. As visualised in Fig. 1a, the Pearson correlation between information density and semantic density for this corpus amounts to $\rho_{Pearson} = 0.711$ ($N_L = 828$, $N = 828$, two-sided non-parametric permutation P-value with 10,000 Monte Carlo permutations, P_{perm}

< 0.001). Since the relationship clearly seems to be non-linear, I also calculated the Spearman correlation that is qualitatively highly comparable, with $\rho_{\text{Spearman}} = 0.791$ ($P_{\text{perm}} < 0.001$). However, Fig. 1b,c clearly demonstrate that both information density and semantic density are strongly correlated with corpus size (in words), with $\rho_{\text{Pearson}} = -0.922$ ($P_{\text{perm}} < 0.001$) and $\rho_{\text{Spearman}} = -0.916$ ($P_{\text{perm}} < 0.001$) for information density and $\rho_{\text{Pearson}} = -0.824$ ($P_{\text{perm}} < 0.001$) and $\rho_{\text{Spearman}} = -0.906$ ($P_{\text{perm}} < 0.001$) for semantic density.

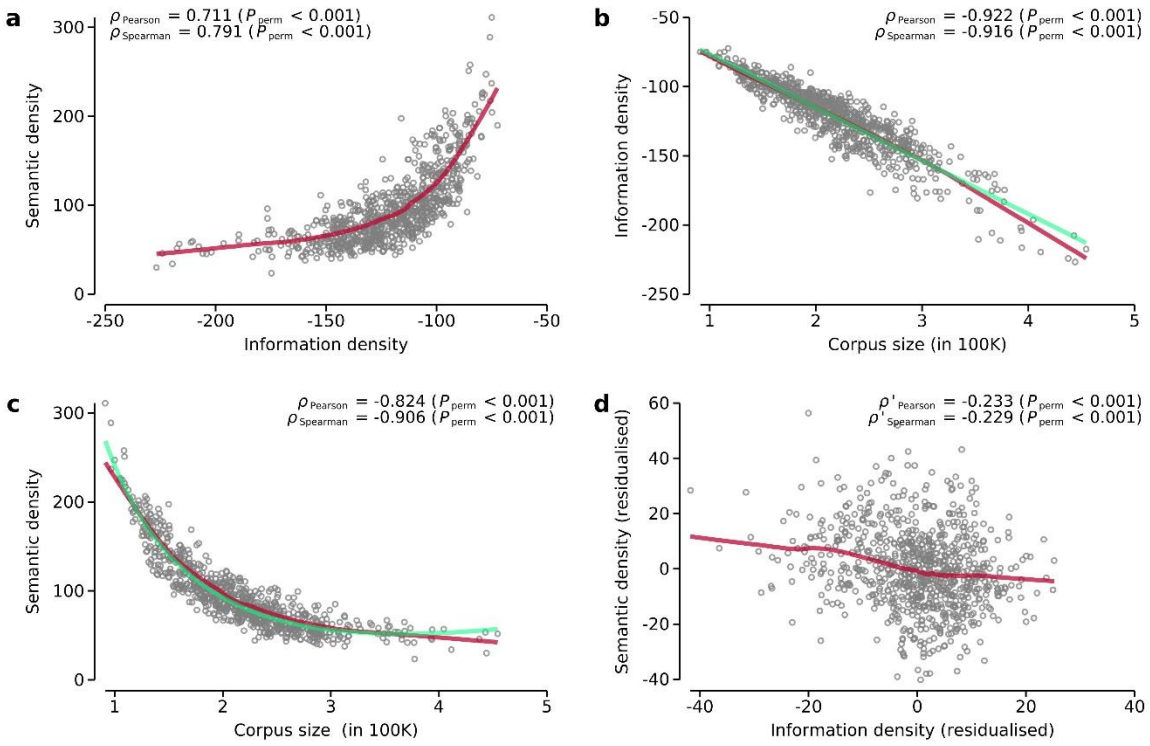


Fig. 1: Relation between information density, semantic density and corpus size (in words) for the Bible_NT corpus for $N_L = 828$ languages and $N = 828$ data points. In all four scatter plots grey circles represent observed data points, while the cranberry line represents a locally weighted scatterplot smoother ("lowess"). The mint line in **b**,**c** represents the parametric prediction described in the text. **a**, Replication of the relationship between information density and semantic density as reported by Aceves and Evans. **b**, Relationship between information density and corpus size (in thousands, K). **c**, Relationship between semantic density and corpus size (in K). **d**, Relationship between information density and semantic density with control for the influence of corpus size.

To test whether this dependence on corpus size affects the results presented by Aceves and Evans, I calculated the correlation between information density and semantic density that would be observed when corpus size were held constant, i.e. partial Pearson ($\rho'_{Pearson}$) and partial

Spearman ($\rho'_{Spearman}$) correlations. To compute $\rho'_{Spearman}$, the variables are first ranked, then the partial correlations are computed in the usual way.[5] To compute $\rho'_{Pearson}$, I first fitted an ordinary least squares (OLS) regression with information density as the outcome and corpus size as the predictor. The mint line in Fig. 1b shows that the model has an excellent fit, with the amount of explained variance, R^2(OLS) = 84.97%. I then fitted an OLS regression with semantic density as the outcome and – to account for the non-linear functional form of the association– the log of corpus size and the log of corpus size squared as predictors. Again the mint line in Fig. 1c reveals that the model has an excellent fit, with R^2(OLS) = 87.33%. I then computed the Pearson correlation between the residuals from both models.

Fig. 1d shows that the influence of corpus size on the observed relationship turns out to be considerable, even reversing the direction of the correlation and thus suggesting a negative relationship between information and semantic density, with $\rho'_{Pearson}$ = -0.233 ($P_{\text{perm}} < 0.001$) and $\rho'_{Spearman}$ = -0.229 ($P_{\text{perm}} < 0.001$). In general, Supplementary Table 1 shows that only in 6 out of 18 investigated corpora, $\rho'_{Spearman}$ is greater than zero at $P < 0.05$ (adjusted for multiple testing).[5]

I proceeded by reanalysing the relationship between information density and semantic density for the whole dataset (N = 1,377). Aceves and Evans report the results (cf. SI Table 5 and SI Fig. 3) of a linear mixed effects regression model (LMM) where semantic density is predicted by fixed effects for information density, corpus size (in words), indicator variables for each corpus category and random intercepts for language family and language nesting observations within languages and languages within families. Fig. 2a replicates this analysis. The effect of language information density, β_{LID} = 1.018 is significant (parametric two-sided P-value, $P_{\text{para}} < 0.001$). To evaluate the model fit, I computed the amount of variance explained by the fixed effect components of the LMM, R^2(LMM), as suggested by ref.[6], eq. 26. For the reported model, R^2(LMM) = 33.74%. Aceves and Evans include the raw value of corpus size as a fixed

covariate. However, Fig. 2b illustrates that the distribution of corpus sizes is significantly right-skewed, violating assumptions of the regression model and adversely affecting the model fit.[7] A common technique for dealing with this problem is to apply a logarithmic transformation.[7] Fig. 2c shows that this approach does reduce the skewness to some extent. However, due to the substantial variation in sizes among the different corpora (see Supplementary Figure 1), the transformed distribution is still skewed. To allow the relationship between corpus size and semantic density to vary across different corpus categories, I fitted a LMM where semantic density is predicted by fixed effects for information density, corpus size (logged), indicator variables for each corpus category, first-order interactions between corpus category and corpus size (logged) and random intercepts for language family and language nesting observations within languages and languages within families. This modified model demonstrates a significantly improved model fit, as reflected by a substantially higher amount of variance explained by the fixed effects, with R^2(LMM) = 57.78%. Fig. 2d visualizes the predicted effect of information density on semantic density for this model. As in the first reanalysis presented above, the direction of the relationship is reversed, with β_{LID} = -0.1321, but does not reach significance at any conventional level (P_{para} = 0.157).

With regard to the reviewer's comment, in my opinion my reanalyses clearly show that the relationships presented by Aceves and Evans are indeed highly dependent on corpus size. Apart from systematic cross-linguistic differences, such as variations in morphological structures and compounding, this finding is especially relevant for the analyses of Aceves and Evans because the corpus data they use vary greatly between languages in terms of text size. For example, within the Subs16 corpus, there are languages with less than 100 available movie subtitles, such as Bengali with 76 and Esperanto with 89, compared to languages such as Swedish with ~27.3K and Italian with ~96.5K subtitles.[8] This further calls into question the overall validity and the interpretation of the results.

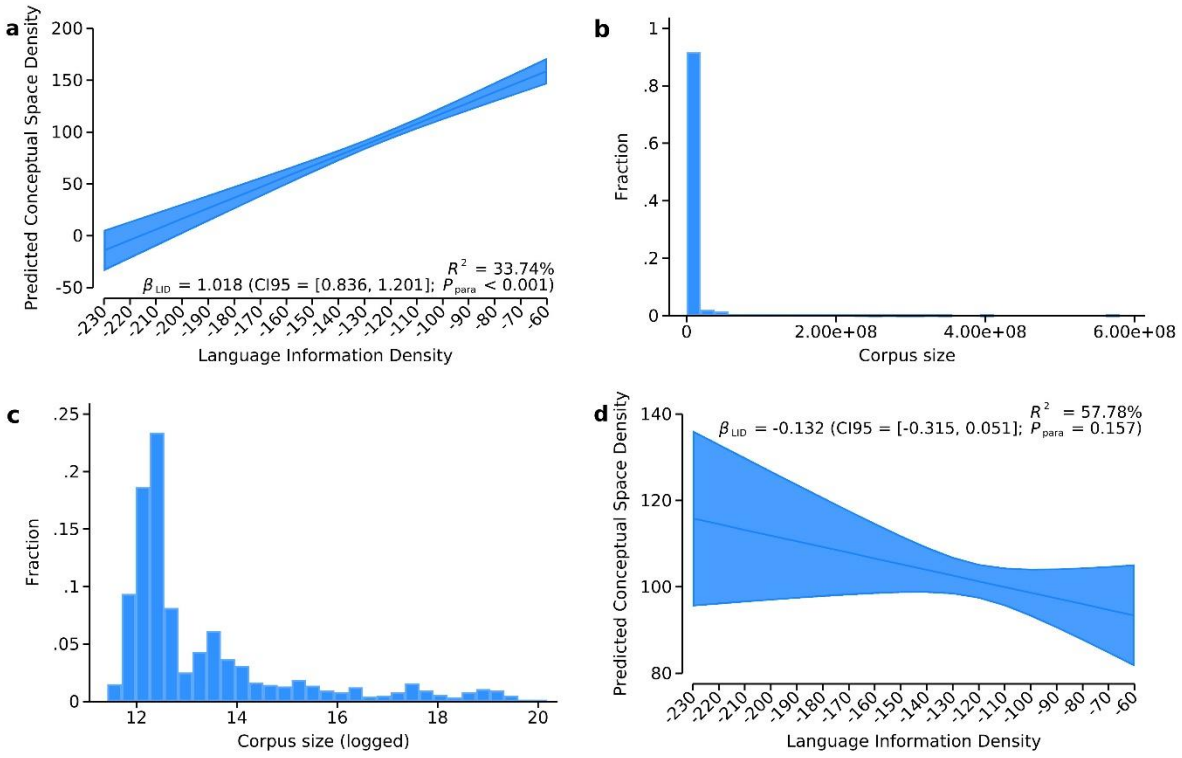


Fig. 2: Relation between information density, semantic density and corpus size (in words) for the for all corpora, N_L = 999 languages and N = 1,377 data points. a, Predictive margins with 95% confidence intervals visualising the relationship between information density and semantic density as reported by Aceves and Evans (cf. SI Fig 3). The LMM with semantic density as the outcome contains fixed effects for information density, corpus size and indicator variables for each corpus category and random intercepts for language family and language. **b**, Histogram visualizing the distribution of corpus size. **c**, Histogram visualizing the distribution of the log of corpus size. **d**, Predictive margins with 95% confidence intervals visualising the relationship between information density and semantic density for a LMM with semantic density as the outcome that contains fixed effects for information density, corpus size (logged), indicator variables for each corpus category, interactions between corpus category and corpus size (logged) and random intercepts for language family and language.

Data availability

The Aceves and Evans dataset is available at https://github.com/peteaceves/Language_Density_and_Communication.

Code availability

Stata (version 18.0) output and code to conduct the reported reanalyses is available at https://osf.io/fmgct/.

Acknowledgments

I thank Pedro Aceves for providing additional data and for a friendly and open discussion. I also thank Peter Meyer and Sascha Wolfer for input and feedback.

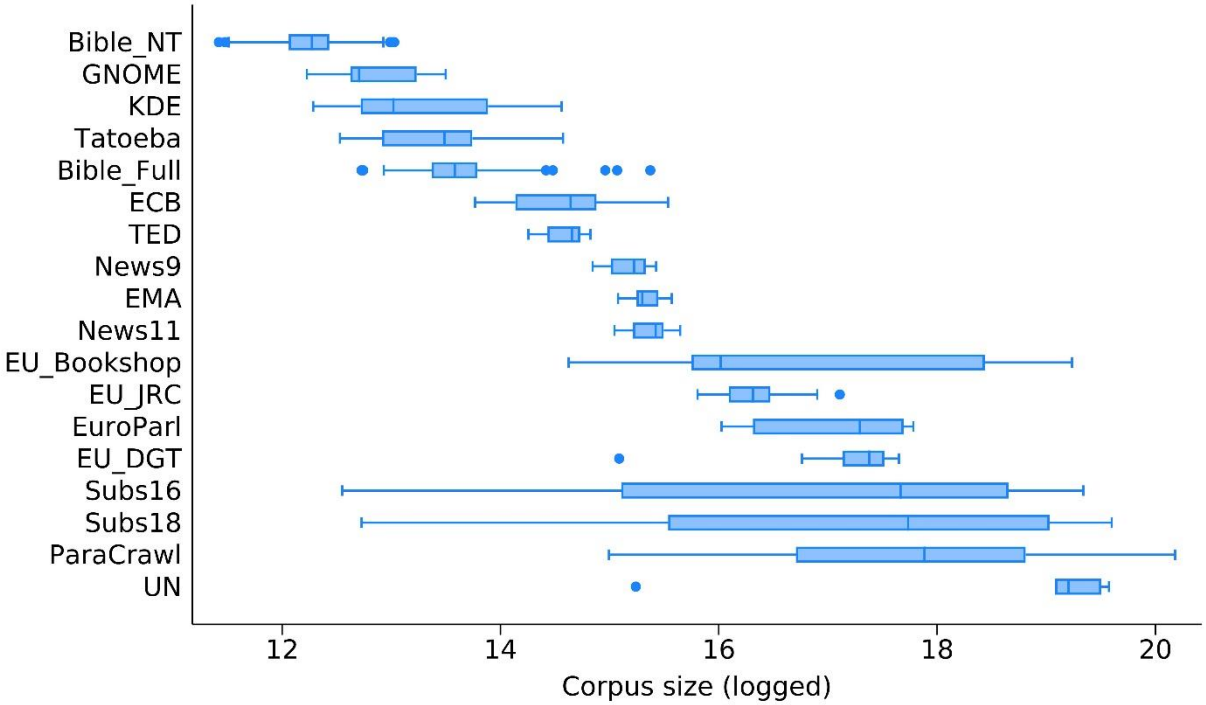
References

1. Aceves, P. & Evans, J. A. Human languages with greater information density have higher communication speed but lower conversation breadth. *Nat. Hum. Behav.* (2024) doi:10.1038/s41562-024-01815-w.
2. Baayen, R. H. *Word Frequency Distributions*. (Kluwer Academic Publishers, Dordrecht, 2001).
3. Tweedie, F. J. & Baayen, R. H. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Comput. Humanit.* **32**, 323–352 (1998).
4. Koplenig, A., Wolfer, S. & Müller-Spitzer, C. Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size. *Entropy* **21**, (2019).
5. Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Ed.* xxviii, 703 (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2003).
6. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).
7. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (Cambridge University Press, New York, 2007).
8. Lison, P. & Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (eds. Calzolari, N. et al.) 923–929 (European Language Resources Association (ELRA), Portorož, Slovenia, 2016).
9. Tukey, J. W. Exploratory Data Analysis. in *The Concise Encyclopedia of Statistics* 192–194 (Springer New York, New York, NY, 2008). doi:10.1007/978-0-387-32833-1_136.

Supplementary Information

Supplementary Table 1 | Partial Spearman correlations between information density and semantic density controlling for corpus size. 1st column: Corpus. 2nd column: Number of cases. 3rd column: Partial Spearman correlation coefficient. 4th column: permutation *P*-value (each with 10,000 Monte Carlo permutations). 5th column: Support for Aceves & Evans is categorised as "yes" if the partial Spearman correlation is positive and significant at the 95%-level and "no" otherwise. To account for multiple testing[5], the Bonferroni-correction is applied i.e. $P_{perm} < 0.05/k$ where k = 18, i.e. the number of tests.

Corpus	*N*	$\rho'_{Spearman}$	P_{perm}	Support for Aceves & Evans?
Bible_Full	222	0.690	0.000	yes
Bible_NT	828	-0.229	0.000	no
ECB	19	0.223	0.364	no
EMA	22	0.477	0.034	no
EU_Bookshop	23	0.777	0.000	yes
EU_DGT	23	0.655	0.002	yes
EU_JRC	16	0.683	0.005	no
EuroParl	21	0.806	0.000	yes
GNOME	10	0.525	0.120	no
KDE	54	0.600	0.000	yes
News11	8	0.883	0.003	no
News9	8	0.588	0.123	no
ParaCrawl	14	0.644	0.013	no
Subs16	45	0.280	0.065	no
Subs18	46	0.443	0.002	yes
TED	14	0.449	0.127	no
Tatoeba	10	0.788	0.005	no
UN	7	-0.393	0.384	no



Supplementary Figure 1 | Visualisation of the distribution of corpus sizes across corpora. Shown are the sizes of the individual language documents (in words, logged) per corpus. The individual plots are sorted by median corpus size in ascending order. Box plot elements are defined as follows: centre line, median; box boundaries, first and third quartiles; whiskers, as defined by Tukey[9]; points, outliers.