Entropy Behaviour upon Dataset Size Update

Louis Estève, Agata Savary, Thomas Lavergne

{first, last}@lisn.upsaclay.fr
Université-Paris-Saclay, LISN–CNRS

Initial idea

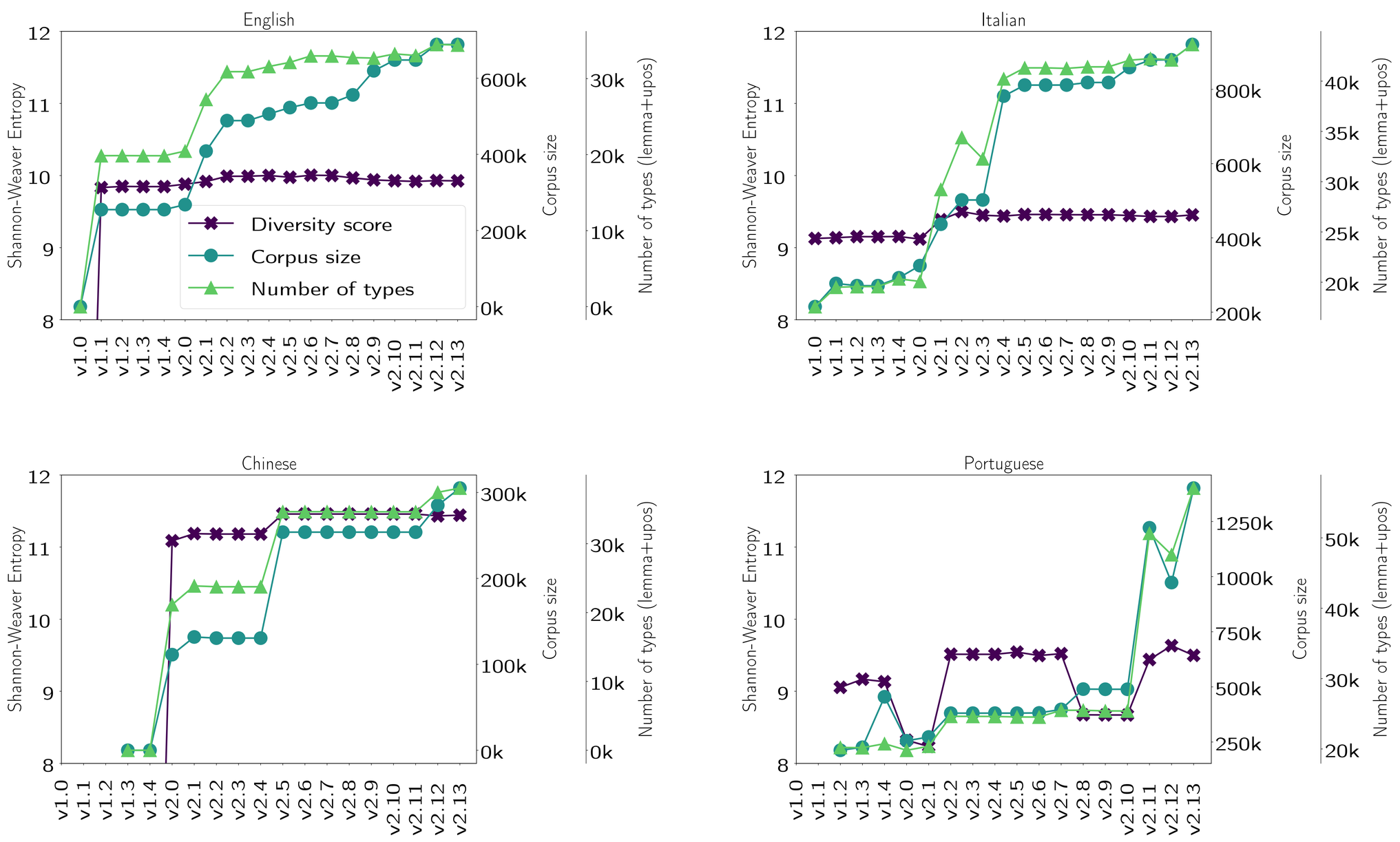
- Datasets often evolve in size across versions
- Before updating a dataset, one may ask: *will this resource increase diversity?*
- Beyond subjective estimates, a formal protocol is desirable

Relevant working groups: WG3, WG4

Question and hypothesis

- Can entropy react to dataset evolution in a way experts deem reasonable diversity-wise?
- Hypothesis: as it measures information, entropy is relevant to estimate the evolution of diversity

Measurements on UD [Nivre et al., 2016]



Analysis

- English: already diverse → close to no increase in H
- Chinese: adding simplified version of data → increase in H
- Italian & Portuguese: more genres → increase in H

Conclusions

- H reacts meaningfully to positive dataset evolution
- H almost ignores non-meaningful dataset evolution
- H reacts to language complexity

Limitations

- H evolves on small scales
- H is not normalised; Shannon Evenness H/H_{max} is a normalised alternative [Morales et al., 2020]

Entropies

[Shannon and Weaver, 1949, Rényi, 1961]

- $H_\alpha = \frac{1}{1-\alpha} \log_b \left(\sum_{i=1}^{n} p_i^\alpha \right)$
- $\lim_{\alpha \to 1} H_\alpha = H = -\sum_{i=1}^{n} p_i \log_b p_i$

References

[Morales et al., 2020] Morales, P. R., et al (2020). Measuring Diversity in Heterogeneous Information Networks. Issue: arXiv:2001.01296 arXiv:2001.01296 [cs, math].

[Nivre et al., 2016] Nivre, J., de Marneffe, et al (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

[Rényi, 1961] Rényi, A. (1961). On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 547–562. University of California Press.

[Shannon and Weaver, 1949] Shannon, C. E. and Weaver, W. (1949). *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Funded by the European Union