Entropy Behaviour upon Dataset Size Update

Louis Estève, Agata Savary, Thomas Lavergne

Entropy Behaviour upon Dataset Size Update

Louis Estève[†] **Agata Savary**[†] **Thomas Lavergne**[†]
[†]Paris-Saclay University, France
Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)

Relevant UniDive working groups: WG3, WG4

1 Introduction

Diversity has the purpose of estimating the overall difference between types (which depending on the task may be NER classes, lemmas, etc.) through *variety*, *balance*, and *disparity* (Lion-Bouton et al., 2022), respectively the number of types, how even their distribution is, and how inherently different they are. Datasets however evolve, often in size, raising the question of whether diversity functions react in a manner experts deem reasonable. We provide in this paper a short multilingual analysis of entropy on the versions of Universal Dependencies (UD, Nivre et al., 2016; 2020); other diversity functions are implemented for future analysis. The choice of entropy is motivated by its importance in information theory, its use as a diversity function itself, and in other diversity functions (Patil and Taillie, 1982; Smith and Wilson, 1996; Mutz, 2022). After presenting entropy (§2) and UD (§3), we experiment on the behaviour of entropy for the lemma+upos pair in four languages across UD versions (§4). We find that entropy reacts meaningfully to dataset evolution.

2 Entropy

Shannon-Weaver entropy H (1949) is the cornerstone of quantifying information; it is defined in a system with n types as

$$H = -\sum_{i=1}^{n} p_i \log_b (p_i) \quad (1)$$

where p_i is the relative proportion of the ith type, and often $b = 2$ to quantify information in *bits*. It is the limiting case of Rényi entropy H_α (1961)

$$H_\alpha = \frac{1}{1-\alpha} \log_b \left(\sum_{i=1}^{n} p_i^\alpha \right) ; \lim_{\alpha \to 1} H_\alpha = H \quad (2)$$

H_α reaches its maximum value $\log_b (n)$ when all types have the same proportion (see appendix). Entropy thus measures balance, and to some extent variety through maximum entropy.

3 Universal Dependencies

At the time of writing, UD has nineteen versions from v1.0 to v2.13, evolving from 10 treebanks over 10 languages to 259 treebanks over 148 languages. It follows the CoNLL-U tabular data format comprising token ID (*ID*), token itself (*form*), lemma (*lemma*), universal part-of-speech tag (*upos*), language-specific part-of-speech tag (*xpos*), morphological features (*feats*), syntactic head (*head*), dependency relation (*deprel*), enhanced dependencies (*deps*).[1] At each subversion, languages can add or withdraw corpora which belong to a number of genres: academic, bible, blog, email, fiction, government, grammar-examples, learner-essays, legal, medical, news, nonfiction, poetry, reviews, social, spoken, web, wiki. One may argue an increase in genres ($\leftrightarrow$ variety) or in evenness between genres ($\leftrightarrow$ balance) is instinctively a positive sign of diversity, to be rewarded by diversity functions. Conversely, a decrease in genres or in their evenness may be seen as a negative sign of diversity, to be penalised.

4 Experiments and Analysis

We see in Figure 1 the evolution of entropy in four languages across UD versions. We first considered studying languages with high/moderate/low amounts of resources, but the need in analysis to have sizeable corpora and substantial updates reduced a lot the options. We will comment on the arrival or withdrawal of corpora, but note that some changes in entropy are caused by substantial corpus reworks, such as Portuguese GSD which increased its annotated lemmas from $\approx 45\%$ in v2.7 to $\approx 71\%$ in v2.8. Note also that we only consider CoNLL-U rows with both lemma and upos defined (often $90 + \%$ of all rows).

The case of English shows that when the main corpora are themselves diverse adding more has a limited effect on entropy (v1.1 $\rightarrow$ v2.13, tokens:$+170.23\%$, lemma+upos:$+73.29\%$,

[1]More details available at https://universaldependencies.org/format.html.

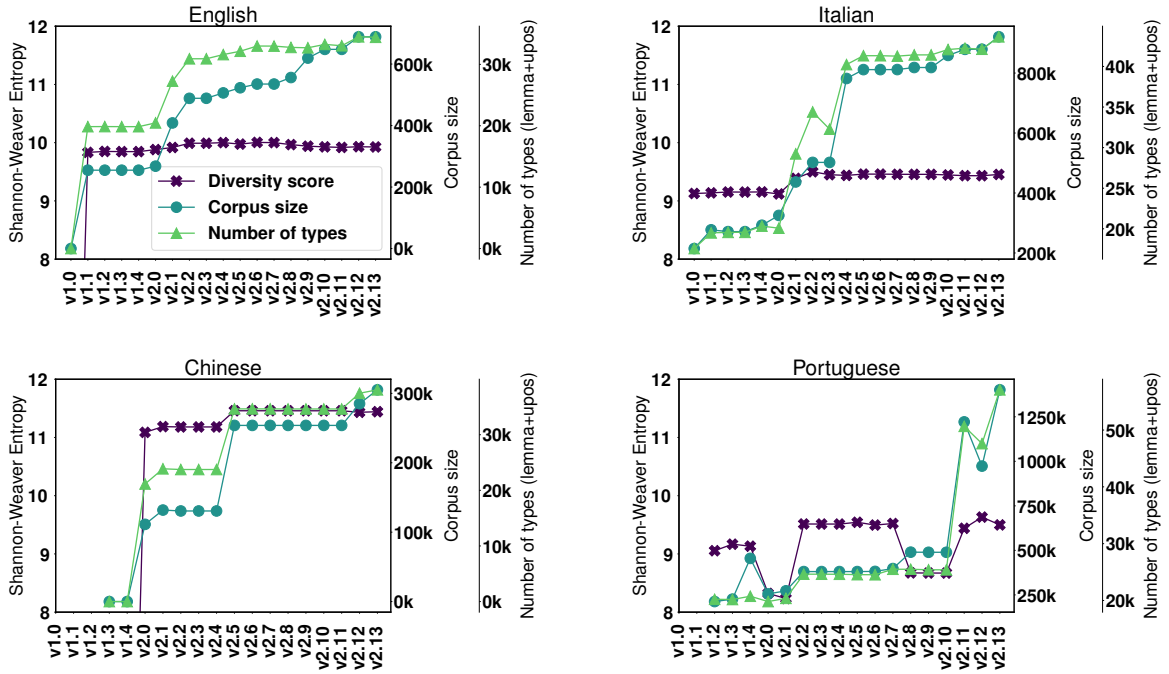

Figure 1: Impact of UD corpus evolution on lemma+upos entropy (in reading order: English, Italian, Chinese, Portuguese). UD_Portuguese-BR appeared in v1.3 and was merged in UD_Portuguese-GSD in v2.2. Corpus size is the number of tokens with both lemma and upos defined.

entropy:+0.94%). EWT the main corpus was present in v1.0, with 5 genres; other corpora have many genres, such as GUM, the second largest corpus, with 9 genres. It should be noted that, despite this genre diversity, the corpora added over time are for many subjectively small; had they been of a similar size as EWT, entropy may have increased more.

For Italian, entropy increases in v2.1 with the arrival (for the most part) of PoSTWITA which brings the new genre social [media] (tokens:+34.30%, lemma+upos:+45.58%, entropy: +2.93%), and diminishes with the arrival of VIT in v2.4 (tokens: +55.73%, lemma+upos: +24.65%, entropy: −0.12%). It may be explained by the fact that one of the two genres of VIT (news) was already present three times in v2.3, combined with a much lower increase in lemma+upos than in v2.1 percentage-wise.

In the case of Chinese, the arrival of GSDSimp in v2.5, a simplified language version of the already present GSD corpus, causes a substantial increase in corpus size with an increase in entropy (tokens: +94.45%, lemma+upos: +45.90%, entropy: +2.48%). Entropy increased despite GSPSimp being of the same genre as GSD (wiki), which may be explained by the difference in language itself.

In a more pronounced manner, Portuguese in v2.11 sees an increase in entropy with the arrival of CINTIL and PetroGold (tokens: +149.63%, lemma+upos: +99.49%, entropy: +8.86%) which increase diversity genre-wise:

v2.10	3 news, 1 blog, 1 wiki
v2.11	v2.10 + 1 news + 1 academic + 1 nonfiction + 1 fiction + 1 grammar-examples

We conclude that entropy is capable of accounting for the instinctive sense of diversity one may have with regards to dataset evolution. For corpus exploration, one may use entropy to gain insights in a uniform manner on the relevance of adding various corpora, for example based on their *forms* if it is prior to annotation. One drawback is that the evolution of entropy, although sensical, is on somewhat small scales, making it hard to properly gauge. $H_{\alpha\neq1}$ and other functions using entropy may improve on this aspect, but the question of whether they keep the sensical behaviour of entropy remains open; existing options consider the number or proportions of types, sometimes with other constants (Smith and Wilson, 1996; Morales et al., 2020).

Acknowledgments

This research was made possible by the "*Plan Blanc*" ("White Plan") doctoral grant from Université Paris-Saclay, France.

A Appendix: maximum entropy

Process showing that for $H_{\alpha \to 1}$, a perfectly even distribution reaches $\log_b(n)$:

$$
\begin{aligned}
H &= -\sum_{i=1}^{n} p_i \log_b(p_i) \\
&\text{if } \forall 1 \leq i \leq n, p_i = \frac{1}{n} \\
H &= -\sum_{i=1}^{n} \frac{1}{n} \log_b\left(\frac{1}{n}\right) \\
H &= -n\left(\frac{1}{n} \log_b\left(\frac{1}{n}\right)\right) \\
H &= -\log_b\left(\frac{1}{n}\right) \\
H &= -\left(\log_b(1) - \log_b(n)\right) \\
H &= -\left(-\log_b(n)\right) \\
H &= \log_b(n)
\end{aligned}
\tag{3}
$$

Process showing that for $H_{\alpha \neq 1}$, a perfectly even distribution reaches $\log_b(n)$:

$$
\begin{aligned}
H_\alpha &= \frac{1}{1-\alpha} \log_b\left(\sum_{i=1}^{n} p_i^\alpha\right) \\
&\text{if } \forall 1 \leq i \leq n, p_i = \frac{1}{n} \\
H_\alpha &= \frac{1}{1-\alpha} \log_b\left(\sum_{i=1}^{n} \left(\frac{1}{n}\right)^\alpha\right) \\
H_\alpha &= \frac{1}{1-\alpha} \log_b\left(n n^{-\alpha}\right) \\
H_\alpha &= \frac{1}{1-\alpha} \log_b\left(n^{1-\alpha}\right) \\
H_\alpha &= \frac{1}{1-\alpha} (1-\alpha) \log_b(n) \\
H_\alpha &= \log_b(n)
\end{aligned}
\tag{4}
$$

References

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating Diversity of Multiword Expressions in Annotated Text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. 2020. Measuring Diversity in Heterogeneous Information Networks. Issue: arXiv:2001.01296 arXiv:2001.01296 [cs, math].

Rüdiger Mutz. 2022. Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? An information-theoretical and statistical estimation approach. *Scientometrics*, 127(12):7397–7414.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

G. P. Patil and C. Taillie. 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*, 77(379):548–561. Number: 379 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Alfréd Rényi. 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.

Claude Elwood Shannon and Warren Weaver. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Publisher: [Nordic Society Oikos, Wiley].