Measuring Grammatical Diversity from Small Corpora: Derivational Entropy Rates, Mean Length of Utterances, and Annotation Invariance

Fermín Moscoso del Prado Martín & Suchir Salhan

Department of Computer Science & Technology, University of Cambridge, U.K. fm611@cst.cam.ac.uk, sas245@cam.ac.uk

Derivational Entropy Rate – A Metric of Syntactic Complexity

Mean Length of Utterance, MLU[G]: A traditional theory-free "proxy" measure of syntactic complexity from unlabelled corpora.

The **Derivational Entropy** of a Grammar, G, is a measure of how uncertain or diverse the derivations are.

Derivational Entropy of a Probabilistic Context-Free Grammar (PCFG)

$$H[G] = -\sum_{t \in T[G]} p_G[t] \log p_G[t]$$

computed using Shannon Entropy *where* $T[G]$ *is the set of parse trees generated by* G *and* $p_G[t]$ *is the probability that a grammar* G *assigns to a tree* t.

Result 1. *Within a PCFG, Derivational Entropy and MLU are directly proportional.* Given this proportionality, we define the **Derivational Entropy Rate** — *the average derivational entropy per unit of string length.*

Derivational Entropy Rate

$$h[G] = \frac{H[G]}{\mathrm{MLU}[G]} = \alpha > 0$$

for constant α of **bits per terminal symbol**

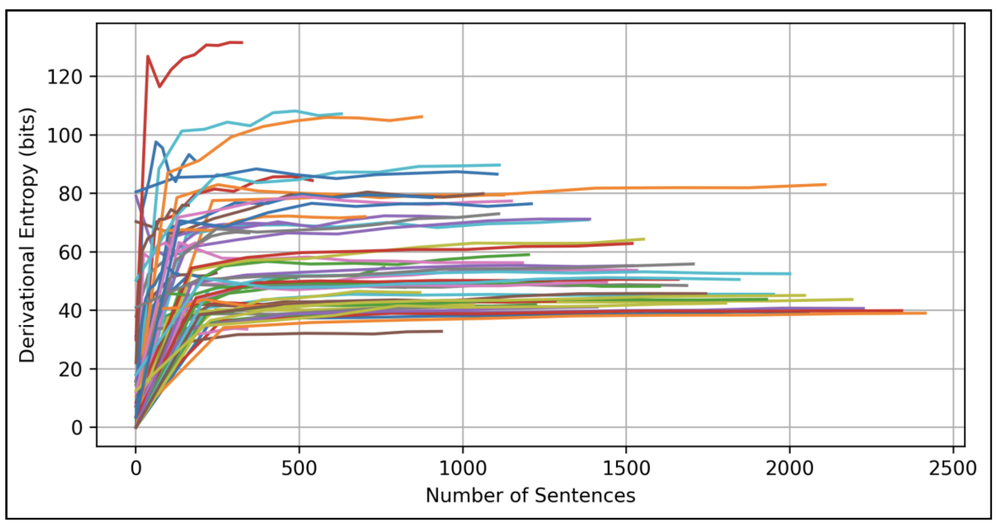
Derivational Entropy Rate indexes the rate at which different grammatical annotation frameworks determine the grammatical complexity of Treebanks.

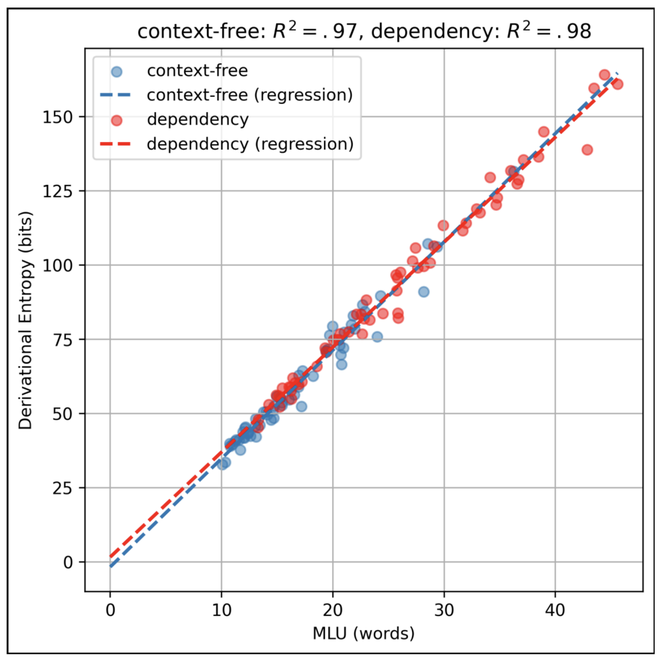
Hypothesis

Derivational Entropy Rate of a given PCFG is **constant across PCFGs** that (1) represent the same or closely related languages AND (2) are annotated using the same grammatical convention.

Smoothed Induced Treebank Entropy (SITE)

SITE is a method for **estimating derivational entropy from Treebanks**, correcting the underestimation of ML entropy estimates. SITE converges to the correct values *quickly and accurately*, even estimating entropy and complexity for **small treebanks** (context-free $\sim$ 100+ and dependency $\sim$ 1000). **Corpus Analyses:** Derivational Entropy Rates are constant across diverse subcorpora.





MLU-Derivational Entropy Correlations

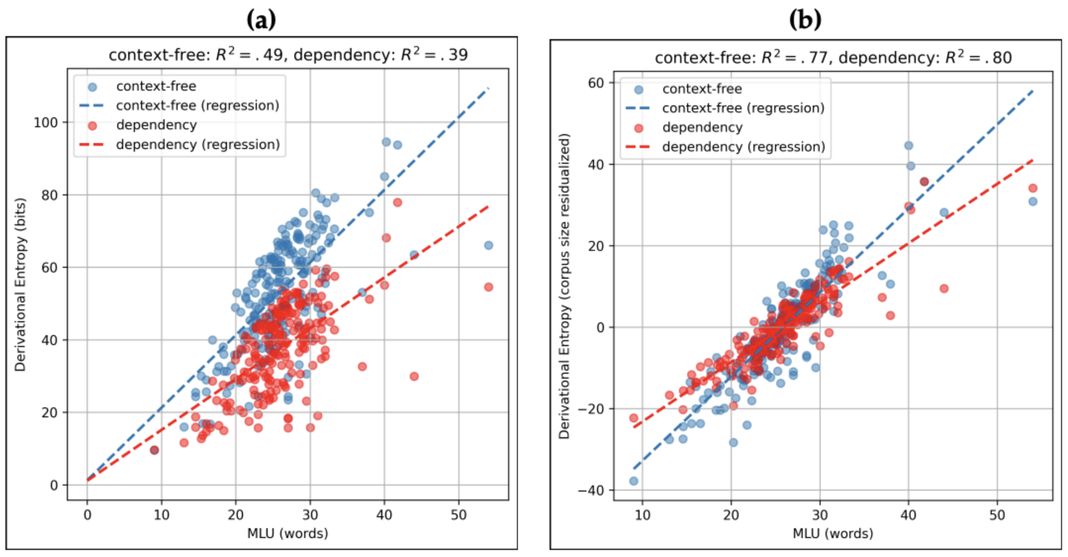


Figure: Relationship between MLU and Derivational Entropy Rate in the samples from Wall Street Journal subsample of the Penn Tree Bank, (a) raw and (b) post-residualisation of log corpus size

Figure: In addition to the Context-Free and Dependency Grammar versions of the Penn Treebank subsample, the original **Icelandic Parsed Historical Corpus (IcePaHC, Rögnvaldsson et al (2012)**, and its Universal Dependencies version. **Top:** Convergence of the SITE derivational entropy estimates. **Bottom:** Relationship between MLU and the derivational entropies for each file in original (blue) + UD (red) IcePaHC.

Key Takeaways

✓MLU is a direct measure of syntactic complexity, not just a proxy

✓Derivational entropy rates are roughly constant across corpora annotated using the same scheme.

✓The constancy of derivational entropy rates can be exploited for obtaining accurate measures of derivational entropy directly from unparsed corpora.