UNIVERSITY OF CAMBRIDGE

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

Fermín Moscoso del Prado Martín[1,2] & Suchir Salhan[1,3]

[1]Department of Computer Science & Technology, University of Cambridge, UK
[2]Jesus College, University of Cambridge, UK
[3]Gonville & Caius College, University of Cambridge, UK

Association for Computational Linguistics Vienna, July 2025

Introduction

- Measuring syntactic complexity in small corpora:
  - many measures available
  - most require substantial corpus annotation
  - Mean Length of Utterance: Classic measure, very robust, but considered a "proxy"
- Derivational Entropy Rate: a robust metric from PCFGs

Derivational Entropy of a Grammar

- For a PCFG G, the derivational entropy is:

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

- Measures uncertainty/diversity of derivations
- Can be computed analytically

Derivational Entropy Rate

- Derivational Entropy Rate: average entropy per unit length

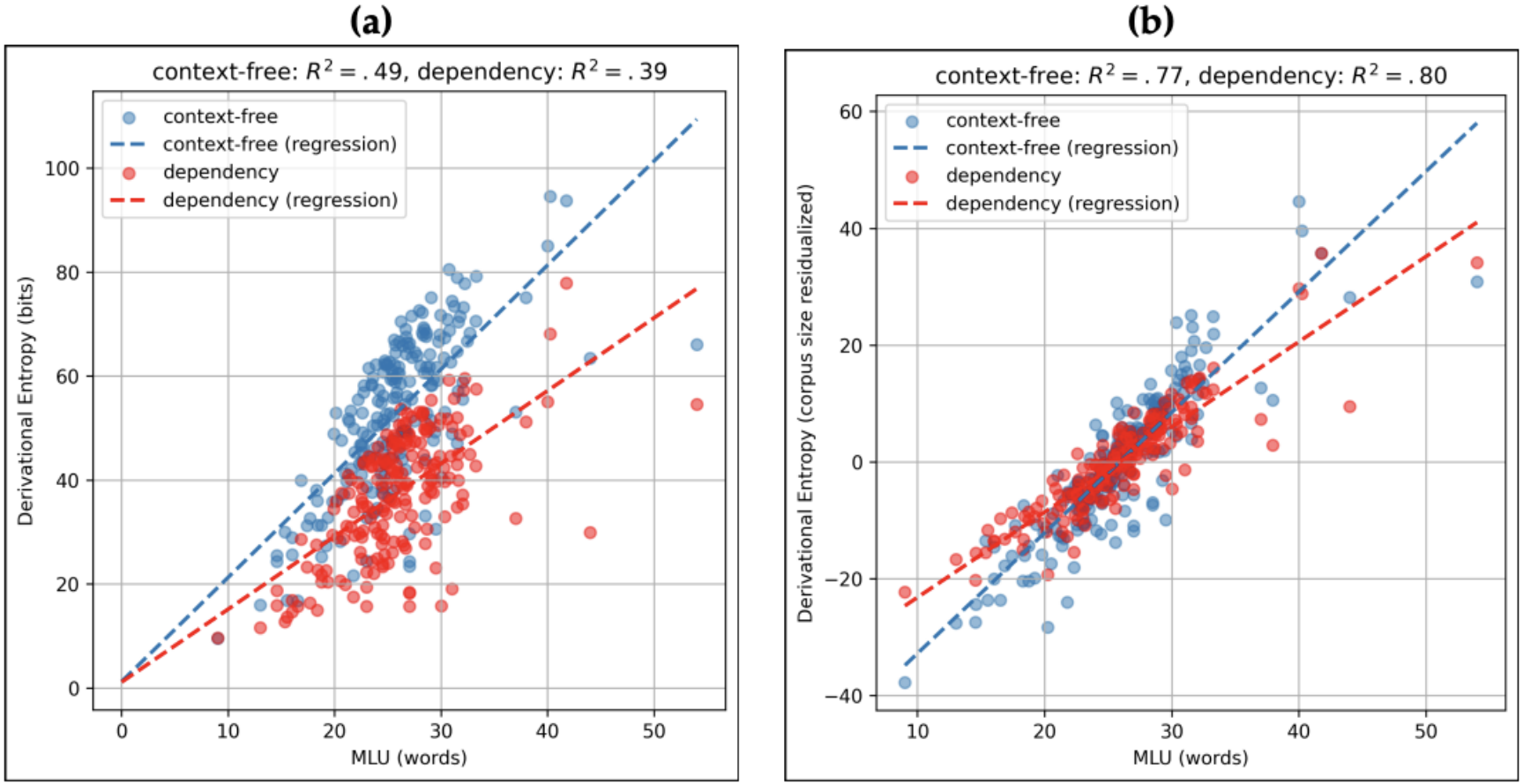
$$h[G] = \frac{H[G]}{MLU[G]} = \alpha > 0$$

- α observed to be constant across corpora

Hypothesis

- Derivational Entropy Rate is constant when:
  1. Same (or related) language
  2. Same grammatical annotation

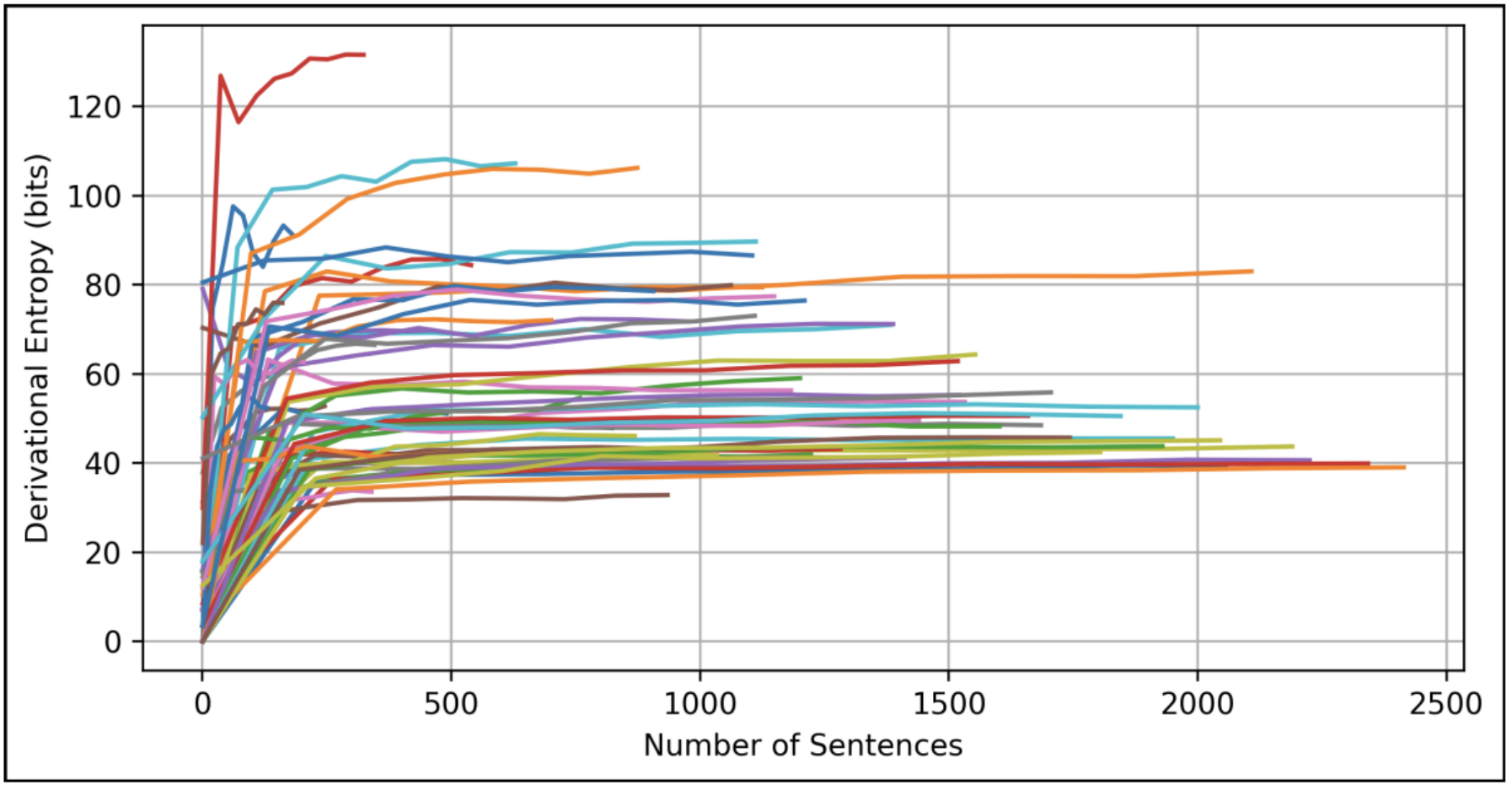
MLU-Derivational Entropy Relationship



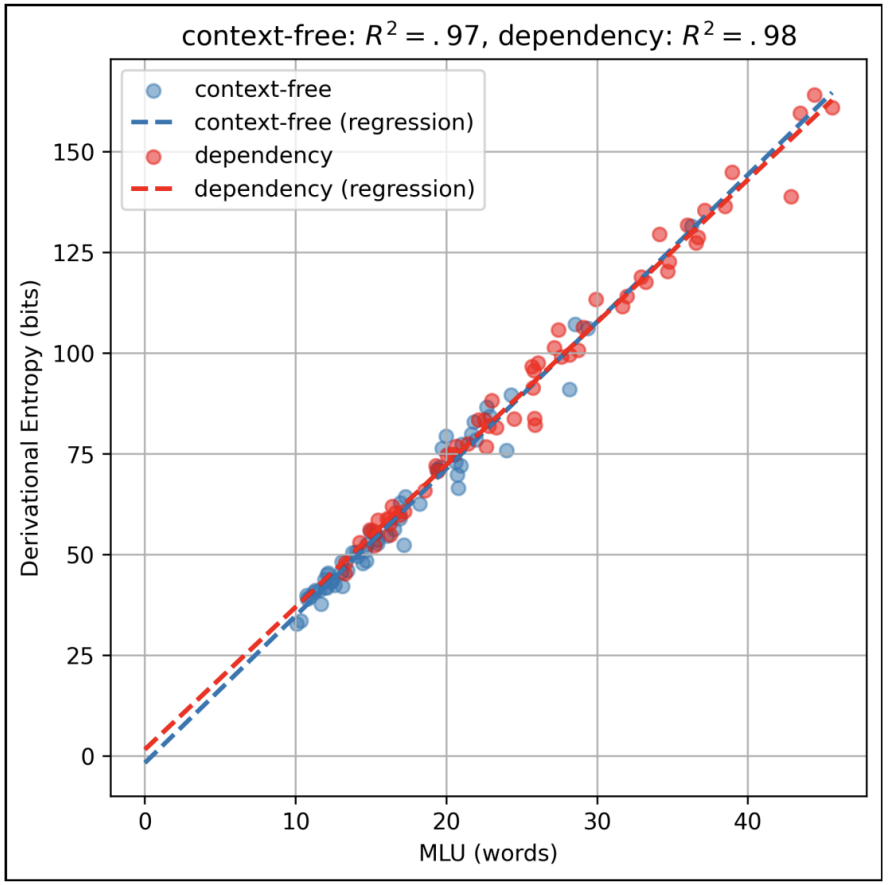
SITE: Smoothed Induced Treebank Entropy

- Corrects ML underestimation of entropy
- Can be used for measuring the derivational entropy of the source that generated a given sample
- Converges very fast, even with small corpora. On realistic PCFGs
  - Context-free: Converges with ∼100 sample sentences
  - Dependency: Converges with ∼1000 sample sentences

SITE: Convergence on Small Corpora (IcePaHC)



SITE: Across two Treebanks



Blue: Original IcePaHC, Red: Universal Dependencies version

Key Takeaways

- MLU is a direct measure of syntactic complexity
- Derivational entropy rates are stable across corpora with consistent annotation schemes
- Enables estimation from unparsed data
- Robust even for small treebanks
- Enables annotation-invariant comparison