

Between order and disorder

An ecological view of lexical complexity measures

Arianna Bienati ^{1,2} Paolo Brasolin ¹

PHRAME: Phraseological Complexity Measures in Learner Italian (closing event)

Perugia, 27th November 2023

¹Eurac Research (Institute for Applied Linguistics), Bolzano, Italy

²Università degli Studi di Modena e Reggio Emilia, Modena, Italy

TABLE OF CONTENTS

1. Background
2. Research questions
3. Method
4. Results
5. Discussion and (preliminary) conclusions

Background

[Complexity is] a matter of the number and variety of an item's constituent elements and of the elaboratedness of their interrelational structure

Rescher, 2020:1

LEXICAL AND PHRASEOLOGICAL COMPLEXITY

De Clercq (2015) suggests **lexical complexity** is a multidimensional construct:

	Meaning	Typical operationalization
Diversity	size of active lexicon	ratio of unique words to total words
Sophistication	difficulty of used words	mean word frequency in reference corpus
Density	information packing	ratio of content words to function words

Paquot (2019) models **phraseological complexity** as a combination of *diversity* and *sophistication* (applied on syntactically dependent units, e.g., V+DO, MOD+N, MOD+V).

A SIMPLE VIEW OF LEXICAL COMPLEXITY

If we look at complexity from a **strictly structural** point of view, we should **discard both density and sophistication** as they do not deal with structure:

*[Using **density** as proxy for complexity] rests on the idea that a certain subset of the lexicon is more complex because it is used by more advanced learners, as there are no clear reasons why, from a purely structural point of view, lexical words should be more or less complex than function words.*

*Similarly, indices of lexical **sophistication**, like the percentage of rare or difficult words, may be valid indicators of development, but they do not directly tap structural complexity; from a structural point of view, a rare word like "tar" is not in itself more complex than a common one like "car".*

Pallotti, 2015:10

DIFFERENT NOTIONS OF COMPLEXITY ...

Classical information theory (Gell-Mann and Lloyd, 1996; Kolmogorov, 1963) offers two very different notions defining the *complexity of a system* that we can borrow:

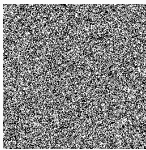
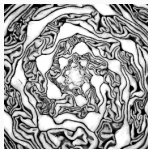
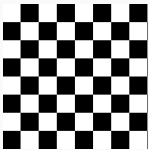
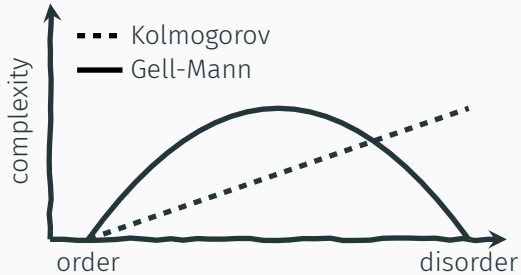
Kolmogorov (algorithmic) complexity	amount of total information
Gell-Mann (effective) complexity	amount of non-random¹ information

Both can be operationalized as the *length of the shortest description of the measured information in a universal description language*. E.g.: Kolmogorov complexity of a string can be defined as the length of the shortest computer program producing it as an output.

None of this is news, and these notions have been employed in typological (Dahl, 2009) and applied linguistics studies (Pallotti, 2015).

¹Defining what *non-random* even means is non-trivial, but we will leave that can of worms closed.

... AND THEIR BEHAVIOURS



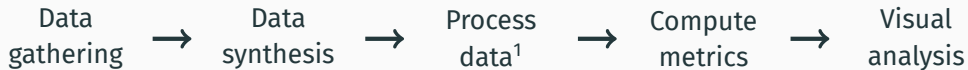
Kolmogorov complexity always increases with disorder. Therefore, it **peaks when there is complete disorder**. **Gell-Mann complexity** increases with disorder only *until randomness starts prevailing*. Therefore, it **peaks when there is rich structure** and vanishes for both simple order and random disorder. This effectively captures the *intuitive* notion of complexity and has been successfully used in other fields already (e.g. in ecology by Parrott, 2010).

Research questions

Which notion of complexity (Kolmogorov vs. Gell-Mann) is measured by the most used lexical diversity indices?

If none of the lexical diversity indices commonly used is measuring Gell-Mann complexity, are there such measures designed in other fields that can be used on texts?

Method



¹This amounts to straightforwardly using Stanford NLP's Stanza to lemmatize the texts for computing metrics.

We picked a few corpora, spanning the spectrum **from experts to learners**:

Corpus	Texts	Tokens	Reference
Eurac Science Blogs	53	79.413	Online
CORIS ²	111	1.359.042	Rossini Favretti et al., 2002
ITACA	635	382.964	Unpublished
LEONIDE	844	81.491	Glaznieks et al., 2022

²We used only the 2014–16 monitor corpus for academic prose.

We generated variants of each corpus, spanning the spectrum **from order to disorder**:

Variant	Description	Example
Repeat 1/16	First 1/16 of the text is repeated	AAAAAAAAAAAAAAAAAAAA
Repeat 1/4	First 1/4 of the text is repeated	ABCAABCAABCAABCA
Original	Text is left intact	ABCABDDEDFDDAGHH
Shuffled	Words are randomly reordered	EBAAEDLIHFGDCBMH
Uniform	Unique words are sampled	DFMACAGCDHMBEDDD

Every text was transformed producing another of the same length.

$$T := \frac{N}{L}$$
$$R := \frac{N}{\sqrt{L}}$$

How diverse are types?

Consider a text made by L tokens of N types ($N \leq L$).

The **type-to-token ratio** T is a simple, well-known quotient. T is regarded as ineffective for comparing texts of different lengths due to the diverging increase rates of L and N .

Guiraud's R is an empirical correction (Guiraud, 1954).

Moving Average TTR is a statistical workaround, and amounts to averaging T on a moving window.

$$H_i := \log\left(\frac{1}{p_i}\right)$$

$$H = \sum_i p_i H_i$$

$$E := \frac{H}{\log(N)}$$

How surprising is the average token?

Let us index the token types of a text with $i = 1, \dots, N$.

Then we denote p_i the probability¹ of observing i .

H_i is the *entropy* (or surprise) associated with the token i .

H is the average token entropy.

H is, in this sense, the **entropy** of the whole text.

H is affected by *lexicon* size, since $\max(H) = \log(N)$ (for $p_i = 1/N$). The **evenness** E is the normalized version of H .

²You might know *probability* as *relative frequency*.

How volatile is the change in surprise with the next token?

$$\Gamma_{ij} := H_j - H_i$$

$$\Gamma = \sum_{ij} p_{ij} \Gamma_{ij}$$

$$\sigma_{\Gamma}^2 = \sum_{ij} p_{ij} \left(\log \frac{p_i}{p_j} \right)^2$$

Γ_{ij} is the *net information gain* from observing token j after i .

Let us denote p_{ij} the probability of observing j after i .

Γ is the average net information gain; it turns out $\Gamma = 0$.

σ_{Γ}^2 however, the mean squared deviation of Γ from its average, need not to be zero and is length-independent by definition.

σ_{Γ}^2 has been used as a complexity metric under the name of **fluctuation complexity** (Bates and Shepard, 1993).

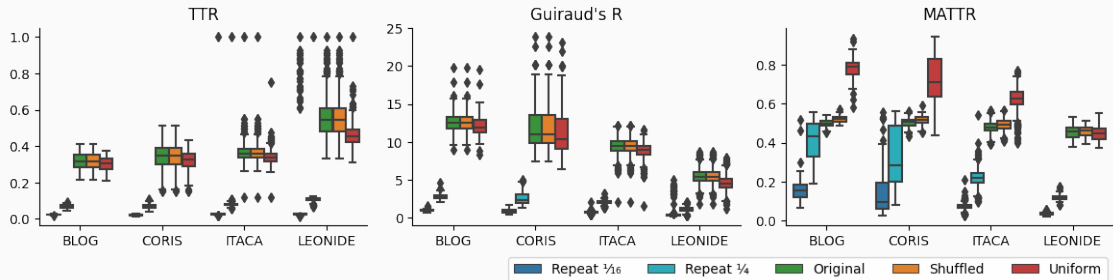
After computing every metric for every text in every variant of every corpus, we facet our dataset and inspect box plots looking for patterns.

We are seeking patterns **across original corpora** along the spectrum from experts to learners, to gauge whether metrics are able to differentiate between them.

We are also seeking patterns **across synthetic corpus variants** along the spectrum from order to disorder, to classify the metrics into Kolmogorov and Gell-Mann complexity measures.

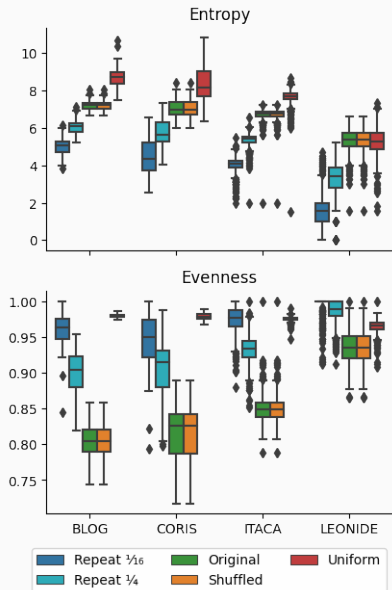
Results

TTR, MATTR AND GUIRAUD'S R



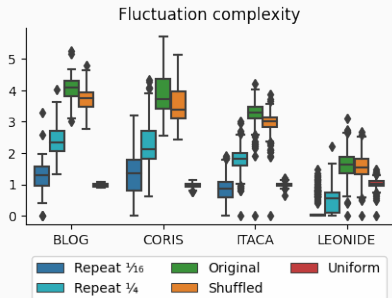
- TTR and R are sensitive to increased order (repetition) and insensitive to increased disorder (no change with shuffling, barely any change on the uniform variant).
- R improves on TTR (higher variance on CORIS, lower values for learner corpora).
- Only MATTR exhibits Kolmogorovian behaviour across variants (except on shorter texts) but is uniform across corpora (probably due to window size).

ENTROPY AND EVENNESS



- Entropy and evenness detect increased order.
- Entropy and evenness partly detect increased disorder: they are insensitive to shuffling but sensitive to the probability changes in the uniform variant.
- Evenness improves on entropy by losing dependence from lexicon size.
- Barring the shuffling behaviour, entropy seems Kolmogorov and evenness seems Gell-Mann.

FLUCTUATION COMPLEXITY



Fluctuation complexity

- is higher on expert corpora,
- is lower on learner corpora,
- is more dispersed on heterogeneous corpora,
- decreases with increasing order,
- decreases with increasing disorder, and
- exhibits markedly Gell-Mann behaviour.

So... do we have a winner?

Discussion and (preliminary) conclusions

DISCUSSION

*[Complexity is] a matter of the **number and variety of an item's constituent elements** and of the **elaboratedness of their interrelational structure**.* Rescher, 2020:1

TTR-based measures, by their very definition, fail to account for the *variety of elements* or *any structure*. They are strongly affected by text-length dependency, and common solutions are either empirical normalization (non-principled) or statistical averaging (losing information). They don't seem to fit either Kolmogorov or Gell-Mann complexity.

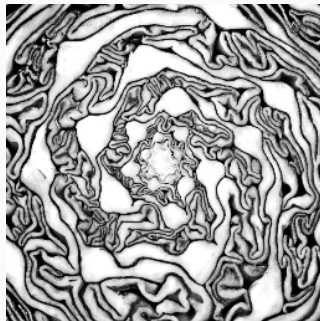
Entropy-based measures also fail to account for *any structure* by definition. However, despite the *bag of words* approach, they are principled notions and they resemble Kolmogorov and Gell-Mann complexities (normalizing by lexicon size).

Fluctuation complexity **satisfies Rescher's definition**: it accounts for the number and variety of elements via their probabilities, and for their interrelational structure via probabilities and net information gains of bigrams. Furthermore, it exhibits markedly Gell-Mann behaviour thus capturing the intuitive notion of effective complexity.

(PRELIMINARY) CONCLUSIONS 1

What are the key takeaways?

- TTR-based measures capture repetition/diversity, not complexity; let's be careful of the construct at hand when using them.
- Entropy-based measures are an improvement but they still see text as a *bag of words*.
- Fluctuation complexity does a better job than all of the above in capturing our intuitive notion of complexity.



Cross section of red cabbage.

(PRELIMINARY) CONCLUSIONS 2

What do we have on our hands?

We seem to have one **well-behaved metric measuring Gell-Mann complexity of texts**.

However, while high scores denote richness of structure, low ones are opaque: are they caused by order or disorder? Computing a Kolmogorov measure could answer that.

So, many open ended questions are open for the next steps:

- Can we find an equally well-behaved Kolmogorov measure?
- How far can we get by complementing two such measures?
- Can we extend our study with new metrics, corpora, and syntheses?

Thank you *very* much!
Any questions?

Arianna Bienati
arianna.bienati@eurac.edu

Paolo Brasolin
paolo.brasolin@gmail.com

- Bates, J. E., & Shepard, H. K. (1993). **Measuring complexity using information fluctuation.** *Physics Letters A*, 172(6), 416–425. [https://doi.org/10.1016/0375-9601\(93\)90232-O](https://doi.org/10.1016/0375-9601(93)90232-O)
- Dahl, Ö. (2009, January 1). **Testing the assumption of complexity invariance: The case of elfdalian and swedish.** In *Language complexity as an evolving variable* (pp. 50–63). Oxford University Press.
- De Clercq, B. (2015). **The development of lexical complexity in second language acquisition: A cross-linguistic study of l2 french and english [Publisher: John Benjamins Publishing Company].** *EUROSLA Yearbook*, 15(1), 69–94. <https://doi.org/10.1075/eurosla.15.03dec>
- Garner, J. (2020). **The cross-sectional development of verb–noun collocations as constructions in L2 writing.** *International Review of Applied Linguistics in Language Teaching*, 60, 909–935. <https://doi.org/10.1515/iral-2019-0169>

- Gell-Mann, M., & Lloyd, S. (1996). Information measures, effective complexity, and total information [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-0526%28199609/10%292%3A1%3C44%3A%3AAID-CPLX10%3E3.0.CO%3B2-X>]. *Complexity*, 2(1), 44–52.
[https://doi.org/10.1002/\(SICI\)1099-0526\(199609/10\)2:1<44::AID-CPLX10>3.0.CO;2-X](https://doi.org/10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X)
- Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of italian, german and english. *International Journal of Learner Corpus Research*, 8(1), 97–120. <https://doi.org/10.1075/ijlcr.21004.gla>
- Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: Essai de méthodologie* [OCLC: 301411769]. Presses universitaires de France. Retrieved November 21, 2023, from <http://catalogue.bnf.fr/ark:/12148/cb37450420s>

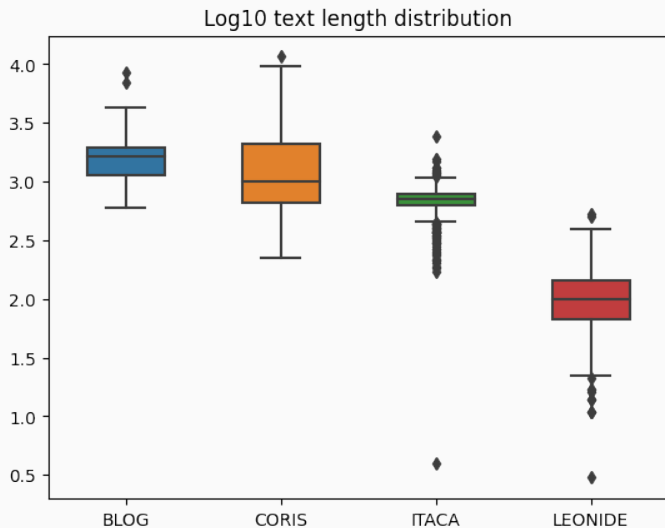
- Kolmogorov, A. N. (1963). **On tables of random numbers** [Publisher: Springer]. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 25(4), 369–376. Retrieved November 20, 2023, from <https://www.jstor.org/stable/25049284>
- Pallotti, G. (2015). **A simple view of linguistic complexity** [Publisher: SAGE Publications Ltd]. *Second Language Research*, 31(1), 117–134. <https://doi.org/10.1177/0267658314536435>
- Paquot, M. (2019). **The phraseological dimension in interlanguage complexity research**. *Second Language Research*, 35, 121–145. <https://doi.org/10.1177/0267658317694221>
- Parrott, L. (2010). **Measuring ecological complexity**. *Ecological Indicators*, 10(6), 1069–1076. <https://doi.org/10.1016/j.ecolind.2010.03.014>
- Proisl, T. (2023, September 30). **Linguistic and stylistic complexity** [original-date: 2017-11-21T08:48:09Z]. Retrieved November 16, 2023, from <https://github.com/tsproisl/textcomplexity>

- Rescher, N. (2020, March 31). ***Complexity: A philosophical overview***. Routledge.
<https://doi.org/10.4324/9780429336591>
- Rossini Favretti, R., Tamburini, F., & De Santis, C. (2002). **CORIS/CODIS: A corpus of written italian based on a defined and a dynamic model**. In A. Wilson, P. Rayson, & A. McEnery (Eds.), *A rainbow of corpora: Corpus linguistics and the languages of the world* (pp. 27–38). Lincom-Europa.
- Rubin, R., Housen, A., & Paquot, M. (2021, August). **5 Phraseological Complexity as an Index of L2 Dutch Writing Proficiency: A Partial Replication Study**. In *5 Phraseological Complexity as an Index of L2 Dutch Writing Proficiency: A Partial Replication Study* (pp. 101–125). Multilingual Matters. <https://doi.org/10.21832/9781788924863-006>
- Vandeweerd, N., Housen, A., & Paquot, M. (2022). **Comparing the longitudinal development of phraseological complexity across oral and written tasks**. *Studies in Second Language Acquisition*, 1–25. <https://doi.org/10.1017/s0272263122000389>

Vandeweerd, N., Housen, A., & Paquot, M. (2021). **Applying phraseological complexity measures to L2 French: A partial replication study*** [Publisher: John Benjamins Publishing Company]. *International Journal of Learner Corpus Research*, 7(2), 197–229.
<https://doi.org/10.1075/ijlcr.20015.van>

Backup slides

WHAT IS THE TEXT LENGTH DISTRIBUTION OF THE CORPORA?



WHICH IS HEAVIER? A KG OF STEEL OR A KG OF FEATHERS?

The Problem with complexity metrics is often framed as one of *text-length dependency*. That is quite misleading, and the confusion arises by conflating *sample size* issues with the lack of *modeling clarity* and distinctions between *intensive/extensive* quantities.

Entropy is extensive. So why doesn't it double when joining two copies of the same text ($H_{1+2} = H_1 = H_2$) or two texts sharing no words ($H_{1+2} = (H_1 + H_2)/2 + 1$)?

That's because **the text is not the system** you're measuring. Tokens are states, the lexicon is the state space, and the text is a trajectory in the state space: **the system is the writer**.

Normalizing to produce intensive quantities is either correct and moot (the size of the system is always *one*) or misled and simply producing new different quantities.

Using longer texts is *not* studying bigger systems: it is taking longer trajectories (bigger samples), to build more accurate *active vocabularies* as more accurate models of the unknowable *mental lexicon* of the writer.

TTR is a property of the sample since it involves text length; it's not measuring anything intrinsic of the model (nor the system).

HOW DID WE CHOOSE THE METRICS?

From the list of lexical complexity measures in Proisl, 2023, we selected those that were also used in phraseological complexity studies:

Metric	Phraseological studies
Guiraud's R	Paquot, 2019, Vandeweerd et al., 2021, Rubin et al., 2021
MATTR	Vandeweerd et al., 2022
Evenness	Garner, 2020