JESLA | Journal of the European Second Language Association

Phraseology meets information theory: Going beyond the *bag-of-words* approach in complexity measures

COLLECTION: MULTIPLE VIEWS ON PHRASEOLOGY IN SECOND LANGUAGE ACQUISITION (PATH: PHRASEOLOGY)

RESEARCH

PAOLO BRASOLIN

ARIANNA BIENATI

*Author affiliations can be found in the back matter of this article

WHITE ROSE UNIVERSITY PRESS
Universities of Leeds, Sheffield & York

ABSTRACT

This article investigates how phraseological diversity measures behave across two different axes of variation – expertise and order – with the aim of determining their ability to discriminate between different levels of expertise in writing as well as their sensitivity to the orderliness of texts. Through a scoping review of phraseological complexity studies, we identify the conceptualizations and operationalizations underlying the phraseological complexity construct. Phraseological complexity relies on modelling interrelationships between words to account for their complexity not in isolation, but in relationship with co-occurring words. After trying extensions of classical lexical diversity measures (e.g., type-token ratio [TTR]-based), we borrow from information theory the measure of information fluctuation complexity, which is able to model the interrelationships between consecutive tokens via token pair frequencies. Using a controlled simulation setting, we apply these measures to four corpora of Italian spanning the spectrum from more expert to less expert writers, as well as synthetic corpora that represent more orderly and disorderly variants of the same texts. Although TTR-based measures computed on bigrams also capture changes in the text structure, fluctuation complexity is the only measure that exhibits a bell-shaped curve, peaking for original texts and decreasing for either orderly or disorderly variants, thus capturing an intuitive notion of complexity.

CORRESPONDING AUTHOR:
Arianna Bienati
Dipartimento di Educazione e Scienze Umane, Università degli studi di Modena e Reggio Emilia, Reggio Emilia, Italy; Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
arianna.bienati@eurac.edu

KEYWORDS:
linguistic complexity; phraseology; information theory; Gell-Mann; learner corpora; simulation

TO CITE THIS ARTICLE:
Brasolin, P., & Bienati, A. (2025). Phraseology meets information theory: Going beyond the *bag-of-words* approach in complexity measures. *Journal of the European Second Language Association*, 9(1), 103–123. DOI: https://doi.org/10.22599/jesla.140

DOI: 10.22599/jesla.140

1. INTRODUCTION

Early L2 research on linguistic complexity has mainly focused on the syntactic and lexical domains (Bulté & Housen, 2012; Kuiken, 2023), with only recent exceptions addressing morphological complexity (Brezina & Pallotti, 2019; Clercq & Housen, 2019). This emphasis on grammar and lexis as separate components aligns with the word-and-rules model of language, which posits that "[t]he vast expressive power of language is made possible by two principles: the arbitrary sound-meaning pairing underlying words, and the discrete combinatorial system underlying grammar" (Pinker, 1998, p. 219). This model, however, overlooks phenomena at the periphery of grammar or lexis – those that do not fit neatly into fully regular and productive syntactic rules, nor fully irregular, non-productive units. Cognitive accounts have particularly criticized the word-and-rules model for falling into the rule-list fallacy (Langacker, 1987), which forces one "to choose between rules and lists" while "there is in reality a third choice [...] both rules and lists" (p. 41).

Constructionist approaches (cf., Goldberg, 2013) follow this third route, describing linguistic systems as composed of *constructions* – form-meaning pairs, present at all levels of description, from morphemes to more abstract phrasal patterns. This perspective does not assume a strict separation between regular patterns processed or produced online governed by syntactic rules and irregular patterns stored in the lexicon. Instead, it posits that abstract patterns, if sufficiently frequent in the input, become gradually entrenched, even if they are fully predictable. These theories, which postulate a continuum of constructions from the most abstract to the most idiosyncratic, better explain phenomena at the lexis-grammar interface, such as complex words, idioms and collocations.

As an important component of both second-language (L2) and first-language (L1) competence, phenomena at the lexis-grammar interface have acquired a paramount position in complexity research in Second Language Acquisition (SLA). This research strand has responded to these theoretical necessities with the construct of phraseological complexity, which accounts for the complexity of word combinations, arising from "the constraints on [a word's] co-occurrence with other words" (Paquot, 2019, p. 123). While the theoretical shifts just described have deeply influenced the construct of phraseological complexity, the measures used to operationalize the construct have not similarly evolved. Some measures of phraseological complexity, and particularly its subcomponent of phraseological diversity, are extensions of lexical diversity measures applied to units larger than a single token. These measures, whether calculated on single or multiple tokens, inherently follow a bag-of-words approach and thus fail to account for relationships between items.

In this article, we present a scoping review of studies on phraseological complexity, aiming to frame its definition and the measures used to operationalize it (see Section 2). Drawing from information theory (Gell-Mann & Lloyd, 1996; Kolmogorov, 1963), in Section 3 we propose using fluctuation complexity (Bates & Shepard, 1993), a measure that captures inter-relationships between co-occurring elements. In Section 4, we test its behavior through a simulation affording highly controlled conditions and compare it to common metrics employed to capture the *diversity* subcomponent of phraseological complexity. The results reported in Section 5 show that extending classical lexical diversity metrics (e.g., type-token ratio [TTR], Guiraud's R [RTTR], moving average type-token ratio [MATTR]) to bigrams is a viable solution to reflect structural variation in texts, though these metrics also exhibit some undesirable behaviors. Only fluctuation complexity combines sensitivity to structural changes with the desirable characteristic of decreasing in highly ordered and disordered configurations. In Section 6, we discuss these findings in relation to existing research on the relationship between phraseological complexity and proficiency, development and writing quality and conclude in Section 7 with study limitations and directions for future work.

2. A SCOPING REVIEW OF PHRASEOLOGICAL COMPLEXITY STUDIES

The present review is defined as a scoping review (Campbell et al., 2023) as it systematically examines a set of studies guided by broader research questions than those typically addressed by systematic reviews. It focuses on research on phraseological complexity in

DOI: 10.22599/jesla.140

L1 or L2 development indexed in the Scopus and Web of Science databases. Studies were last retrieved on 3 May 2024 using the keyword "phraseological complexity" in the search interface (reference list available at https://doi.org/10.5281/zenodo.15078024). During manual screening, four papers were excluded for using the term only in funding acknowledgements, without addressing the concept substantively. Additionally, one paper was not retrievable due to being behind a paywall and inaccessible with any of the authors' institutional credentials. The review is thus based on a total of 18 papers, systematically analyzed for

- definitions of linguistic complexity
- definitions and operationalizations of phraseological complexity

For empirical studies, we also recorded

- research questions (and hypothesis)
- data
- results

An Excel file summarizing these elements, along with a narrative synthesis of key findings, is available as supplementary material at https://doi.org/10.5281/zenodo.15078024. In this section, we focus on definitions of (phraseological) complexity and provide a detailed overview of the quantitative indices used to measure phraseological complexity, including their interpretation and potential limitations.

2.1. LINGUISTIC COMPLEXITY

Although researchers acknowledge that linguistic complexity in L2 research lacks a clear, agreed-upon definition (e.g., Paquot, 2019; Vandeweerd et al., 2023) and that the available definitions are far from well conceptualized (e.g., Esfandiari & Ahmadi, 2023; Zhang & Ouyang, 2023), the recognition of its importance is ubiquitous in the field. It plays a crucial role as an index of L2 proficiency, along with accuracy and fluency (e.g., Jiang et al., 2023; Vandeweerd et al., 2021). In what follows, we present the range of definitions found in the reviewed papers, grouped by whether they include the concept of *inter-dependency*.

Milani et al. (2021) and Zhang & Ouyang (2023) cite Rescher's (1998) definition of complexity, which states that "complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of their inter-relational structure, be it organizational or operational" (p. 1). This definition highlights three main components: number, variety and the elaborateness of the inter-relational structure among elements that make up the system. While number and variety are relatively straightforward to conceptualize and implement, elaborateness is harder to define. It requires a clear understanding of what qualifies as structure in the linguistic theory of reference as well as in the linguistic domain (e.g., syntax, lexicon, morphology vs. morphosyntax) under investigation. Despite these challenges, inter-dependency is a core aspect of complexity definitions in several reviewed studies. For example, Massip-Bonet et al. (2019), cited by Forti (2020, p. 29), describe complex systems as comprising "a series of elements that are deeply interwoven and interdependent in their functioning". Similarly, Pallotti (2015) defines complexity as "the complexity directly arising from the number of linguistic elements and their interrelationships" (p. 117), a definition adopted by Kim & Crossley (2023) to ground their empirical work.

A different take on complexity comes from Ortega (2003), who defines "syntactic complexity (also called syntactic maturity or linguistic complexity)" as "the range of forms that surface in language productions and the degree of sophistication of such forms" (p. 492). This definition emphasizes number, variety and sophistication of linguistic forms and omits inter-dependency. It underpins the conceptualization and operationalization of the phraseological complexity construct in studies such as Paquot (2019), Rubin et al. (2021) (a replication on L2 Dutch of Paquot 2019) and Vandeweerd et al. (2023).

2.2. PHRASEOLOGICAL COMPLEXITY: DEFINITION AND MEASURES

All reviewed papers except one (Hu et al., 2022) adopt Paquot's (2019, p. 124) definition of phraseological complexity: "the range of phraseological units that surface in language

DOI: 10.22599/jesla.140

production and the degree of sophistication of such phraseological units". This definition builds on Ortega's (2003) definition of linguistic complexity and frames phraseological complexity as comprising two sub-components: diversity and sophistication.

Hu et al. (2022), by contrast, treat diversity, sophistication and complexity as separate constructs. However, to the best of our understanding, they do not clarify what is meant by phraseological complexity (as opposed to diversity and sophistication), nor which construct their operationalization – a ratio of specific word combination types to total combinations – is intended to measure.

Despite the widespread adoption of Paquot's (2019) definition, the indices used to operationalize diversity and sophistication vary widely, both in formulae and in methods for extracting phraseological units. We do not engage here with the definitions or extraction methods of phraseological units; it suffices to note that operationalizations span all n-grams, n-grams with specific part-of-speech tags, and dependency relations. Instead, we focus on the indices used to measure diversity and sophistication, acknowledging the broader variability in how phraseological units are defined and extracted.

Tables 1 and 2 list the metrics used in the reviewed works to measure the diversity and sophistication of phraseological units. Focusing first on sophistication, three main strategies can be identified: association scores, frequency bands and collocation lists. The most common approach (used in 12 out of 18 studies) relies on mean association scores such as mean Pointwise Mutual Information (PMI), LogDice or t-scores. These scores, calculated using a reference corpus, indicate the strength of association of a given unit in the learner text. The rationale behind this calculation is that more strongly associated combinations reflect greater phraseological sophistication than those at or below the chance level. In particular, "by promoting the relatively less frequent and more semantically complex word pairs in learner production, [P]MI taps into the phraseological complexity, and more particularly the phraseological sophistication of learner writing" (Paquot, 2019, p. 125). PMI scores can also be grouped into bands, with varying proportions of units falling into each.

Table 1 List of phraseological complexity metrics measuring the diversity component.

DIVERSITY MEASURE	REFERENCES
TTR	Guzzi & Alonso Ramos (2023), Yin et al. (2024)
RTTR	Paquot (2019), Rubin et al. (2021), Vandeweerd et al. (2021), Hu et al. (2022), Jiang et al. (2023), Esfandiari & Ahmadi (2023), Biondi et al. (2023)
MATTR	Vandeweerd et al. (2023)
HD-D	Hu et al. (2022)

Table 2 List of phraseological complexity metrics measuring the sophistication component.

SOPHISTICATION MEASURE	REFERENCES
proportion of academic collocations	Paquot (2019), Rubin et al. (2021)
proportion of collocations in bands based on PMI	Paquot (2018), Rubin et al. (2021), Vandeweerd et al. (2021), Guzzi & Alonso Ramos (2023)
proportion of collocations bands based on normalized counts and Log-Likelihood	Guzzi & Alonso Ramos (2023)
mean PMI	Paquot (2018), Paquot (2019), Paquot et al. (2021), Rubin et al. (2021), Rubin (2021), Vandeweerd et al. (2021), Kim & Crossley (2023), Vandeweerd et al. (2023, calculated over 100-word moving windows), Jiang et al. (2023), Esfandiari & Ahmadi (2023)
mean t-score	Kim & Crossley (2023)
mean LogDice	Biondi et al. (2023)
proportion of low-frequency word combinations with positive PMI	Hu et al. (2022)
mean frequency	Zhang & Ouyang (2023), Yin et al. (2024, log-transformed)
mean range (dispersion)	Zhang & Ouyang (2023)

DOI: 10.22599/jesla.140

A similar band-based approach is found in Guzzi & Alonso Ramos (2023), who assigned collocations to sophistication bands based on their normalized frequencies in two corpora: one representing general Spanish, the other academic Spanish. Since the more sophisticated words in one register are the ones specific to that register (Read, 2000), the authors considered collocations that are more frequent in academic Spanish to be more sophisticated. Log-likelihood scores were used to identify such collocations and assign them to higher bands, while those more frequent in general Spanish were placed in lower bands.

A third strategy uses academic collocations lists, such as that presented in Ackermann & Chen (2013). Sophistication is then calculated as the proportion of collocations from the list found in the text. As Esfandiari & Ahmadi (2023) observe, all these approaches rely on external reference corpora or lists, which have the advantage of linking individual writing samples (in a Saussurean sense, *la parole*) to the wider representation of the language system (*la langue*). However, they also risk variability, as results may drastically change depending on the reference corpus used, sometimes with unintended consequences (e.g., Bottini & Le Foll, 2024).

Diversity measures, by contrast, are usually text-internal (i.e., calculated solely from the language sample under investigation, usually based on token and type counts). In seven of the reviewed studies, diversity is measured with RTTR (Guiraud, 1954), either on word forms or lemmas. RTTR adjusts the basic type-token ratio by dividing the number of types by the square root of tokens, thus partially mitigating the effects of Herdan–Heaps's law, which predicts diminishing returns in type growth as token count increases. However, RTTR is still somewhat sensitive to text length. For studies where text length could significantly impact results, alternative metrics like the moving average type-token ratio (MATTR) or HD-D are used. All cited measures are sensitive to the number (tokens) and variety (types) of elements but do not explicitly model inter-relations. Any relational aspect is implied only if the units themselves – such as n-grams or dependency-linked elements – are assumed to be interrelated. Originally designed for single words, diversity metrics treat phraseological units as unanalyzed chunks, ignoring internal structure and varying strengths of association between components. As bag-of-words approaches, they overlook the order of appearance and the broader distribution of unit elements within the text. It remains an open question whether simply analyzing larger units than single tokens is enough to meaningfully capture the structural organization of a text.

3. MOTIVATION AND RESEARCH QUESTIONS

While the complexity studies reviewed here rely on descriptive theoretical frameworks prominent in linguistics (e.g., generative vs. cognitive and constructionist models), a few exceptions in both theoretical (Futrell & Hahn, 2022; Gibson et al., 2019) and applied linguistics (Alzahrani, 2024; Ehret & Szmrecsanyi, 2016; Wang et al., 2022) adopt representation-independent approaches to complexity. These alternative accounts draw on notions and quantities from information theory, which provides "notions of complexity by quantifying intrinsic lower bounds on resource requirements for transforming or storing information" (Futrell & Hahn, 2022, p. 12). Information-theoretic concepts and measures are neutral with regards to linguistic theories, since they work with minimal linguistic assumptions about the code to be quantified.

Classical information theory offers two distinct notions in defining the complexity of a system: *Kolmogorov complexity* and *Gell-Mann complexity*. Both can be operationalized as the length of the shortest description of the measured information in a universal description language. However, they differ significantly in what constitutes information worth measuring. While Kolmogorov complexity measures the total amount of information, including randomness, Gell-Mann complexity measures only non-random information. To exemplify, we refer to the widely used and cited example provided by Dahl (2009, p. 51).

> Suppose we have three strings of characters, *hahaha*, *byebye*, and *pardon*. Although these all consist of six characters, they differ in that the two first strings can in fact be represented in a more compact way, e.g. as 3 × *ha* and 2 × *bye*, whereas there is no way of compressing the string *pardon* in a similar way. We might therefore say that *hahaha* is the least complex string, since it can be reduced to four characters, while *byebye* takes minimally five and *pardon* six characters. As applied to strings, this notion of complexity, which is sometimes called "Kolmogorov complexity" or

DOI: 10.22599/jesla.140

"algorithmic information content", comes out as an inverse of compressibility: the most complex string is one which cannot be compressed at all. However, this would mean that maximal complexity would be represented by a random combination of characters, since such a combination cannot be compressed in any way. An alternative would be what Gell-Mann (1994) calls "effective complexity", which differs from Kolmogorov complexity in that it does not measure the length of the description of an object as a whole, but rather the length of the description of the "set of regularities" or structured patterns that it contains. A random string of characters, such as "w509mf0wr6435217ro0l71734", will have maximal Kolmogorov complexity (the string is its own shortest description), but no effective complexity since it contains no structured patterns. This corresponds better to an intuitive understanding of the notion of complexity.

This distinction between measured information leads to different expected behaviors, visually illustrated in Figure 1. Kolmogorov complexity peaks with a completely unpredictable system, as each unit's value must be specified to have a complete description. In contrast, Gell-Mann complexity peaks when there is rich structure, thus when there is neither absolute predictability nor absolute unpredictability. As noted by Dahl (2009), for linguistic data, where interest lies in governing rules rather than randomness, Gell-Mann complexity appears more suitable. However, defining non-random information remains a non-trivial problem.

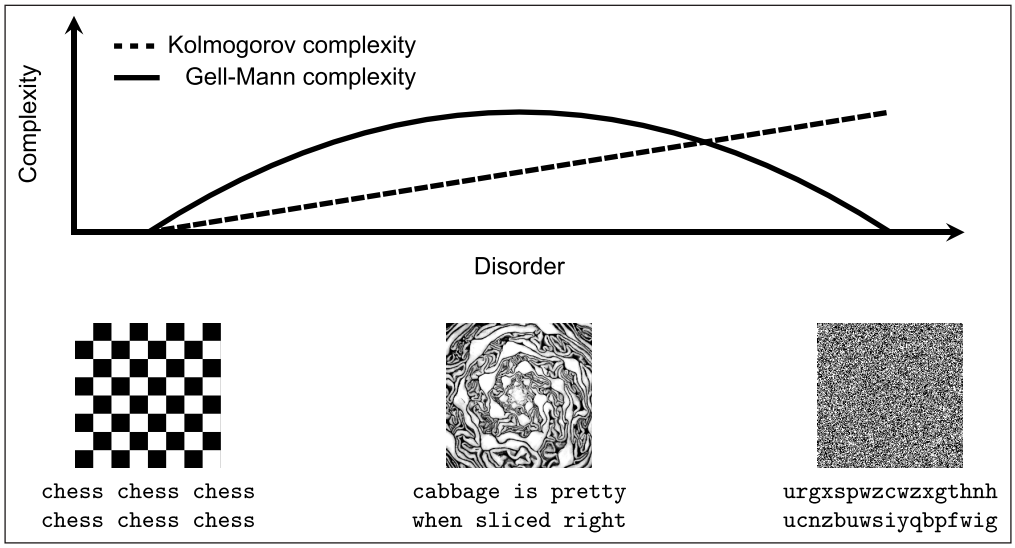

Figure 1 Qualitative behavior of Kolmogorov and Gell-Mann complexity measures in relation to disorder, with illustrative examples at minimum, maximum and intermediate levels.

Examining complexity from a theory-neutral perspective offers several advantages. It can bridge applied and theoretical linguistics, while also connecting the field to disciplines like computer science, physics and ecology. This perspective facilitates the transfer of methodological tools from the hard sciences into linguistics, information-theoretic measures being a prime example. Notably, PMI, originally derived from information theory, has been widely adopted in linguistics as an association measure, including in phraseological complexity research.

Based on these considerations, we aim to address the following research questions:

1. Are there complexity measures derived from information theory that can model syntagmatic ties within texts and exhibit Gell-Mann behavior?
2. What kind of behavior do widely used phraseological diversity indices exhibit?

4. METHOD

To answer these research questions, we evaluate multiple measures on texts varying across two axes:

- the spectrum of expertise, from L1 experts to L2 learners,
- the spectrum of order, from trivial repetition (maximal order) to complete randomness (minimal order).

DOI: 10.22599/jesla.140

To span the spectrum of expertise, we selected (sub)corpora with well-characterized literacy levels and language backgrounds. To span the spectrum of order, we generated synthetic corpora from the natural ones by introducing either order or disorder in incremental steps. The resulting values were then aggregated to characterize the measures by comparing between them and across spectra.

Our resources are publicly available, as follows:

- source code of the data pipeline for subcorpora extraction, tokenization, lemmatization and measure computation: https://doi.org/10.5281/zenodo.15077976;
- dataset of the computed measures: https://doi.org/10.5281/zenodo.15066681;
- computational notebook for analysis: https://doi.org/10.5281/zenodo.15077999.

Running the pipeline may be difficult, as some input corpora were unpublished at the time of writing or not publicly available. To support reproducibility, the output dataset with all computed measures is provided, allowing the analysis to be replicated from the processed data. We also encourage its use for conceptual replication studies on other datasets.

4.1. DATA

The datasets used for the analysis, as well their synthetically generated variants, are presented in the following sections.

4.1.1. Acquisition

Table 3 presents the five subcorpora used in the analysis. They are all small to medium-sized subsections of four Italian corpora that represent academic writing, on a cline from less expert to more expert L1 and L2 writers. Their selection follows the common practice of validating measures against expected linguistic patterns across developmental stages or proficiency levels.

Table 3 Subcorpora employed for the analysis.

CORPUS	SUBCORPUS	LITERACY LEVEL	TEXTS	TOKENS	REFERENCE
LEONIDE_IT	Arg. text, L2	6th-8th grade	166	13556	Glaznieks et al. (2022)
LEONIDE_IT	Arg. text, L1	6th-8th grade	143	17292	Glaznieks et al. (2022)
Kolipsi-2_IT	Arg. text, L2	12th grade	883	173785	Glaznieks et al. (2023)
ITACA	Arg. text, L1	12th grade	495	353259	Bienati et al. (forth.)
PEC	Academic prose	MA-PhD	240	1209843	Spina (2014)

LEONIDE participants are the least expert writers of the cohort. They are lower secondary students who have just started their training in so-called "academic language", a variety characterized by structural peculiarities that differentiate it from everyday language (Mastrantonio, 2021). In educational contexts, this variety is deemed more appropriate to convey subject-specific contents and to perform argumentative or deliberative acts. Training of academic writing occurs mostly through the production of argumentative essays. To ensure consistency in text type across datasets, we selected only the argumentative texts from both the LEONIDE and Kolipsi-2 corpora, excluding picture stories.

For lower- and upper-secondary students, we were able to pair the L1 data (LEONIDE and ITACA) with L2 data (LEONIDE and Kolipsi-2), in order to gain insights into how phraseological diversity measures comparatively behave in texts written by language learners that are differently exposed to academic varieties of Italian. All these corpora were collected in South Tyrol (Italy), a multilingual area where German and Italian are co-official languages with separate monolingual schooling systems. Students from any language background can attend either German or Italian schools, meaning that in German schools there might be students with Italian as one of the family languages and vice versa. We use both the students' declared first languages and the language of education as sampling criteria. The L1 data is thus composed of students who are enrolled in South Tyrolean schools with Italian as a medium of instruction and declare Italian as one of their first languages. The L2 data contains texts written by students

DOI: 10.22599/jesla.140

who are enrolled in South Tyrolean schools with German as a medium of instruction (where Italian is taught as an L2) and do not declare Italian as one of their first languages. These sampling criteria, which admittedly exclude some profiles, are chosen in order to control as much as possible for the potential variation in exposure to the Italian language, which is to be expected in the context in which the data was collected.

On the upper end of the expertise spectrum, we conceptualized expert writers as university students (earning their BA or MA) and research staff from Italian universities. A sample of their writing is collected in the academic subsection of the Perugia Corpus (PEC). These texts not only represent the written productions of cognitively mature individuals who have been exposed to and trained in academic language for more than 11 years but are also texts that have usually been revised by experts in the discipline, a procedure that improves both the quality of the linguistic form and the content.

There are major mean text length differences between texts in the different corpora, with the LEONIDE texts being the shortest and the PEC texts being the longest. The text length distributions of the subcorpora employed in the study are compared in Figure 2.

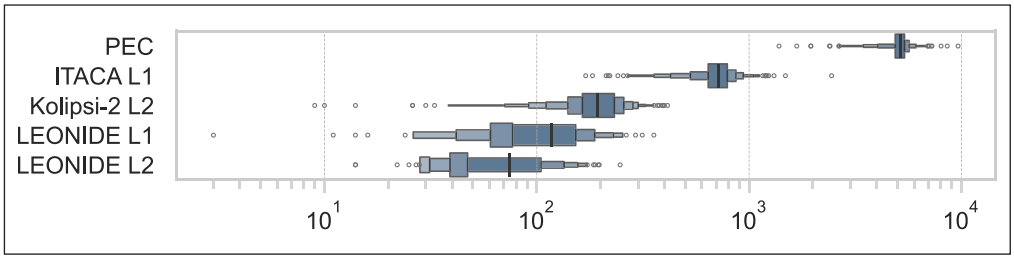

Figure 2 Text length distribution across subcorpora.

Note. Dots are outliers and central boxes represent central quartiles; outer boxes correspond to further percentiles, each half as wide as the previous ($\frac{1}{8}$, $\frac{1}{16}$, ...) box. Since their height is proportional to the data density, the overall shape of the distributions is showcased.

As the aim is to comparatively assess the strengths and weaknesses of widely used and innovative complexity measures, our study design leverages these differences to test whether and how much complexity measures co-vary with the text length and the extent to which the problem can be solved by statistical (e.g., moving averages) or empirical (e.g., derived from corpus-based studies) corrections.

4.1.2. Preparation

All selected texts were prepared by using the Stanza v. 1.7.0 pipeline (Qi et al., 2020) to tokenize and lemmatize them. Punctuation was preserved and multi-word tokens were split. The result was saved in CoNLL-U format (see https://universaldependencies.org/format.html) and used in all subsequent steps. Lemmas are thus our basic unit of analysis and we define an n-gram as a tuple of *n* adjacent lemmas.

4.1.3. Synthesis

To span the spectrum of order, we generated synthetic corpora from the original ones by transforming texts introducing either order or disorder in incremental steps. A *repetition* algorithm was applied to texts to introduce order and was as follows: split the text into chunks of given length (discarding the remainder), then randomly choose one and repeat it to fill the original length (discarding the excess). The process was repeated with incrementally smaller chunks to introduce more repetitions and thus more order; given a text of length *L*, the chunk sizes were $max([L/n],1)$ for $n = 2, 4, 8, 16, 32, 64, 128, 256, 512$.

A *shuffling* algorithm was applied to texts to introduce disorder and was as follows: split the text into chunks of given length (keeping the remainder), then randomly shuffle them and concatenate them into a new text with the original length. The process was repeated with incrementally smaller chunks to introduce more randomness and thus more disorder; the chunk sizes were *n*, for $n = 512, 256, 128, 64, 32, 16, 8, 4, 2, 1$.

Finally, a *resampling* algorithm was applied to texts to introduce even further disorder and was as follows: get the list of token types (unique words) from the original text, then randomly sample from this list (with replacement) to create a new text of the same length as the original.

DOI: 10.22599/jesla.140

4.2. METRICS

Before reviewing all metrics chosen for this study, we will establish some basic conventions. All metrics were computed on both 1-grams and 2-grams. The computations are formally identical and in the current section we will therefore denote either choice as simply *token*. The meaning of all symbols denoting the basic quantities we need for any given text are as follows:

- L is the number of tokens in the text;
- N is the number of (unique) token types in the text;
- Token types are labelled with numbers 1, ..., L;
- n_i is the number of occurrences of tokens of type i;
- $p_i = \frac{n_i}{L}$ is the relative frequency of tokens of type i;
- n_{ij} is the number of occurrences of token pairs of type (i, j);
- $p_{ij} = \frac{n_{ij}}{L}$ is the relative frequency of token pairs of type (i, j).

4.2.1. Type-token ratio

The *type-token ratio* (TTR) is a well-known measure requiring little explanation. It is calculated using the following formula:

$$\frac{N}{L}$$

Its main limitation is that it is unsuitable to compare texts of substantially different lengths. This limitation can be understood in the context of the Herdan-Heaps' empirical law (Corral & Font-Clos, 2017; Serra-Peralta et al., 2021), stating that N and L are related by a power law, that is, $N \simeq \beta L^{\alpha}$ for some constants $0 < \beta$ and $0 < \alpha \leq 1$. The definition of TTR conflates vocabulary size with vocabulary richness, ignoring the fact that the fulfillment of Herdan-Heaps' law suggests vocabulary size grows non-linearly as a function of text length.

4.2.2. Guiraud's index

Guiraud's index (RTTR) is an empirical correction to the type-token ratio based on a corpus study (Guiraud, 1954) and is computed as:

$$\frac{N}{\sqrt{L}}$$

In the context of Herdan-Heaps' empirical law (Corral & Font-Clos, 2017; Serra-Peralta et al., 2021), this correction can be understood as assuming the universality of value $\alpha = \frac{1}{2}$; RTTR itself is then the β parameter, as $\frac{N}{\sqrt{L}} = \frac{\beta L^{\alpha}}{\sqrt{L}} = \frac{\beta\sqrt{L}}{\sqrt{L}} = \beta$.

4.2.3. Moving average of the type-token ratio (MATTR)

MATTR is a statistical correction to the type-token ratio computed as its average on a moving window. While MATTR circumvents the issues TTR has with text length, it does have its own limitations. The main difficulty of using MATTR is the necessity to choose a window size: This is not straightforward and also significantly impacts the results. For this reason we computed it using a variety of window sizes to compare the outcomes: 64, 128, 256 and 512 tokens. When window sizes exceeded the length of a text, we excluded it from the computation of averages; we explicitly note whenever this exclusion affected our analysis.

4.2.4. Entropy

Information Theory centers on the concept of information as a measure of uncertainty. The key idea is that the *information content* of an event E (e.g., a coin toss resulting in heads) is inversely related to its probability of occurrence p and is quantified as $I_E = -log_2 p_E$. Due to the properties of the logarithm, I_E is zero if $p_E = 1$ and tends to infinity as p_E approaches zero; in plain English, this means that information content is minimal for certain events and it quickly increases for rarer events, thus explaining why *surprisal* is a common alternative name for information content. The unit of measure of information is the *bit* (e.g., 1 bit is exactly the information content of a fair coin toss, since $-log_2 \frac{1}{2} = 1$).

DOI: 10.22599/jesla.140

Given all possible outcomes O of a trial (e.g., a coin toss has O = {*heads, tails*}), an interesting quantity is the expected amount of information conveyed by observing the outcome of a trial. This is the *average information content* (AIC) over all possible outcomes and is commonly called *entropy*: $H = \sum_{E \in O} p_E I_E$.

Turning to token types (i.e., the classes of unique tokens, as opposed to specific instances) as the outcomes of observing a token in a text, all notions above immediately translate to texts. The *information content* – or rather, surprisal – associated with observing a token of type i is $I_i = -log_2 p_i$, and the *entropy* of a text is $H = \sum_i^N p_i I_i$, or equivalently

$$H = -\sum_i^N p_i log_2 p_i$$

4.2.5. Fluctuation complexity

Entropy relies on a bag-of-words model and cannot account for the order of words in a text. To capture aspects of sequential information in language, we need to consider more sophisticated measures.

The next simplest information-theoretical quantity that can be considered to account for this limitation is the *net information gain* associated to observing event F after event E, simply defined as the difference in their information content $\Gamma_{EF} = I_F - I_E$. Applying this notion to texts, it is tempting to compute the average net information gain over token pairs $\Gamma = \sum_{i,j=1}^N p_{ij} \Gamma_{ij}$ and its mean square deviation from the average $\sigma_\Gamma^2 = \sum_{i,j=1}^N p_{ij} (\Gamma_{ij} - \Gamma)^2$. It can be shown that that Γ is always zero. Proving this fact requires mildly sophisticated dynamical systems theory that will not be discussed here (see Bates 2020 for a tutorial and Wackerbauer et al. (1994) for a detailed review).

The square root of σ_Γ^2 is called the *information fluctuation complexity* (IFC) and quantifies the volatility of the net information gain between tokens:

$$\sigma_\Gamma = \sqrt{\sum_{ij}^N p_{ij} \left(log_2 \frac{p_i}{p_j} \right)^2}$$

Since σ_Γ is the standard deviation of Γ, the meaning of IFC can be said to be the volatility of the change in surprisal between token pairs.

To further improve our understanding of IFC and how it relates to features at different scales, we compute its moving average using a variety of window sizes: 64, 128, 256 and 512 tokens.

4.3. VISUALIZATION

Table 4 summarizes all measures computed in this study with their abbreviations.

Table 4 Full list of the computed measures along with their abbreviations.

LABEL	MEASURE
TTR	Type-token ratio
RTTR	Guiraud's R
MATTR (w)	Moving average of TTR with window size w = 64, 128, 256, 512
AIC	Entropy (average information content)
IFC	Information fluctuation complexity
MAIFC (w)	Moving average of IFC with window size w = 64, 128, 256, 512

Computing all measures (for both 1-grams and 2-grams) on all texts produces a dataset of considerable size, just below 2 million data points. To effectively analyze this dataset, we group data by measure and subcorpus (either natural or synthetic) and visualize the resulting distributions using boxplots. Each boxplot illustrates how a measure varies along the spectrum from order to disorder. The blue boxplots on the left represent order, while those in red and brown on the right represent disorder. The original texts appear in the center as green boxplots. Boxplots are organized in grids, allowing comparisons between measures and subcorpora,

DOI: 10.22599/jesla.140

effectively displaying the full spectrum of expertise. Each boxplot shows the median of a dataset as a line, then two central quartiles as boxes adjacent to it and extend whiskers on either side for a further $\frac{3}{2}$ of the interquartile range. The remaining data points are considered as outliers and shown as dots.

Looking only at the graphs, some differences can be difficult to capture. In these cases, we support our claims with effect sizes, that is, the magnitude of difference between two groups. We choose Cliff's delta ($\boldsymbol{\delta}$, Cliff, 1993), computed using the *effsize* package (Torchiano, 2020), as it does not require any assumptions about the shape or spread of the distributions of the two groups being compared. Its values intuitively range from -1 to 1, "with 0 indicating stochastic equality of the two groups. 1 indicates that one group shows complete stochastic dominance over the other group, and a value of –1 indicates the complete stochastic domination of the other group" (Mangiafico, 2016, p. 247). Estimates for comparisons on the expertise (original vs. original) and order spectrum (original vs. permutation) are provided in the supplementary materials. Here we only report a qualitative assessment of the magnitude of effect size (i.e., if the effect is negligible [$|\boldsymbol{\delta}|$<0.147], small [$|\boldsymbol{\delta}|$<0.33], medium [$|\boldsymbol{\delta}|$<0.474] or large [$|\boldsymbol{\delta}|\geq$0.474]), following Romano et al. (2006) in its interpretation.

5. RESULTS

We now systematically examine the behaviors of all measures computed in this study across the spectra of order and expertise. As discussed, we expect that Kolmogorov measures will increase monotonically with increasing disorder, while Gell-Mann measures will be concave and decrease towards either extremum (see Figure 1).

5.1. MEASURES ON 1-GRAMS

We start by illustrating the results on measures computed on 1-grams.

5.1.1. TTR-based measures

Figure 3 shows the distribution for TTR, RTTR and MATTR (64). TTR and RTTR are clearly sensitive to increased order in the form of repetition and show an exponential decrease towards zero, which reflects the rate at which the chunk sizes get smaller. This is expected: Both TTR and RTTR are proportional to *N*, which is at least halved with each step. By comparison, both measures are clearly insensitive to increased disorder in the form of shuffling and show a perfectly flat graph. This is also expected: Both follow the bag-of-words model and are therefore by definition insensitive to the order of tokens.

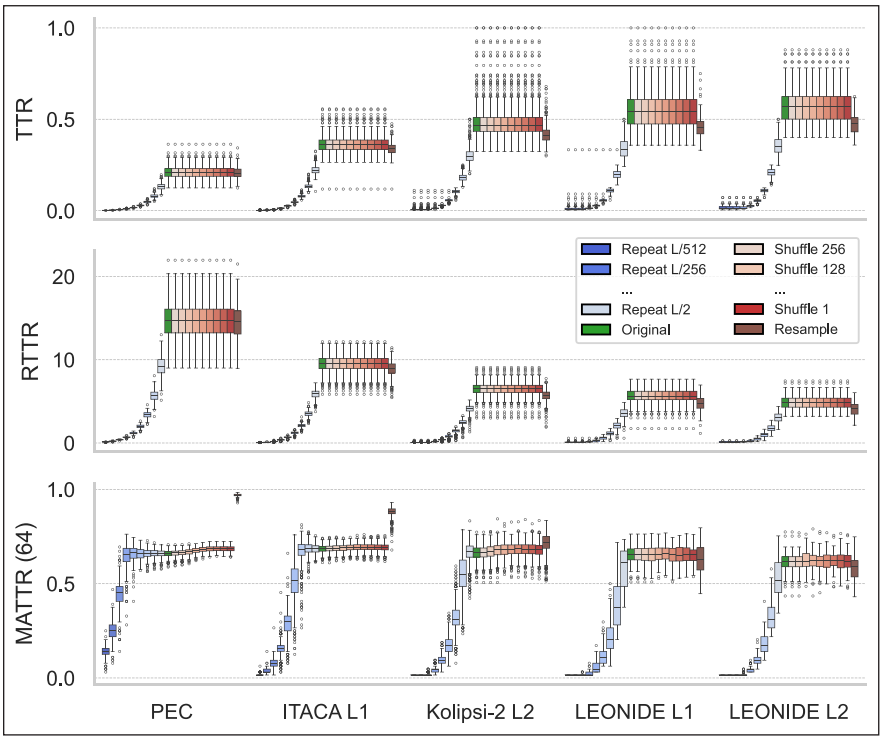

Figure 3 TTR-based measure distributions for 1-grams across all subcorpus variants.

DOI: 10.22599/jesla.140

The measures are mildly sensitive to increased disorder in the form of uniform resampling. This is due to the fact that resampling the tokens with uniform probability effectively reduces *N*, as there is no guarantee of sampling each one of the original token types at least once. The effect is more pronounced (with larger effect sizes) on corpora with shorter texts which have fewer samplings and a relatively larger *L* (compared to *N*).

Figure 3 shows the distribution of MATTR for window size 64 only, a choice forced by the short text lengths of LEONIDE (we discuss other window sizes shortly). MATTR is clearly sensitive to increased order just like TTR, although the behavior is more complex. While it still tends to zero, it is flat until the size of the repeated chunk becomes bigger than the window size. Since the chunk sizes depend on the text length, this causes different curves on each subcorpus and a substantial amount of dispersion (as text length varies within each subcorpus).

MATTR is very mildly sensitive to increased disorder in the form of shuffling; for subcorpora with longer text lengths, a small increase can be noticed when the size of the shuffled chunk becomes equal to or smaller than the window size. Following this, the curve flattens again. By comparison, it is clearly sensitive to increased disorder in the form of resampling; for subcorpora with longer text lengths, we observe very high values due to *N* being vastly bigger than the window size. As the text lengths (and consequently *N*) decrease and become closer to the window size, a behavior analogous to TTR is recovered.

On the order spectrum, neither TTR, RTTR or MATTR exhibits an overall behavior that is convincingly either Kolmogorov or Gell-Mann. On the expertise spectrum, we observe that RTTR is an improvement over TTR as it has lower values for learner corpora, with large effect sizes. MATTR, at least for this choice of window size, shows a decrease in the magnitude of the differences between them. We will now consider other differences, detectable only with a closer inspection across window sizes.

Figure 4 shows the distribution of MATTR for a wide range of window sizes. The usual "tradeoff" of performing moving averages is immediately apparent: Larger window sizes increase sensitivity but force the exclusion of shorter texts (see the missing boxplots in Figure 4), while smaller window sizes include more texts at the cost of flattening out differences and decreasing sensitivity. Notwithstanding this flattening, L1 corpora show slightly higher values compared to their L2 counterparts. The same is true for older L1 and L2 speakers compared to younger speakers. PEC, by contrast, shows slightly lower values than ITACA L1, especially for smaller windows (MATTR [window size 512]: ITACA L1 > PEC, $\boldsymbol{\delta}$ = –0.152, small; MATTR [window size 256]: $\boldsymbol{\delta}$ = 0.075, negligible; MATTR [window size 128]: $\boldsymbol{\delta}$ = 0.356, medium; MATTR [window size 64]:

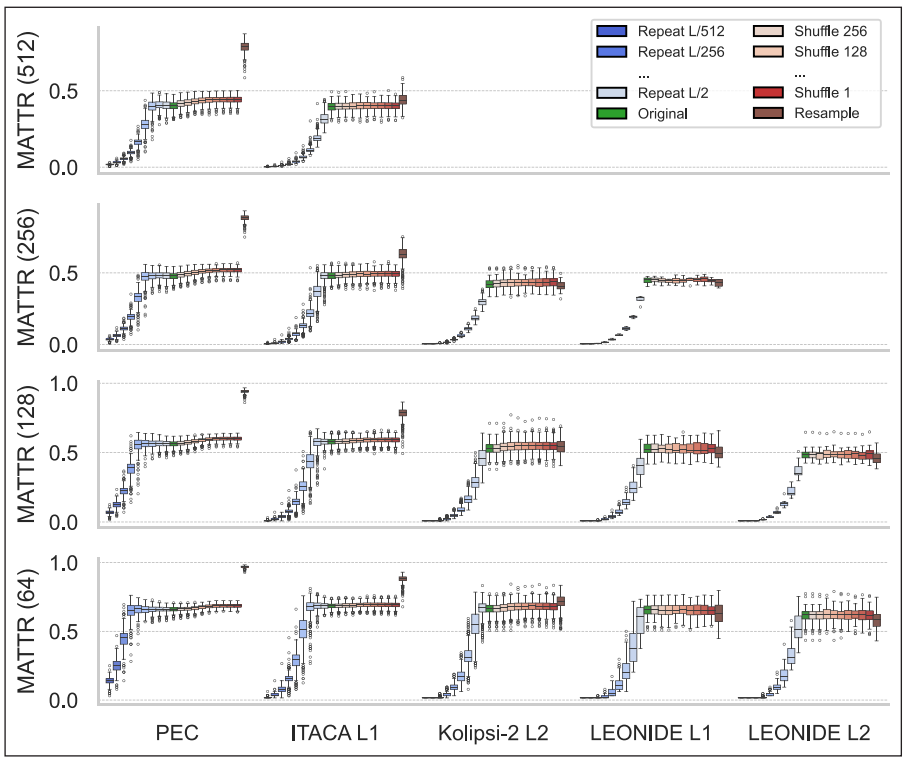

Figure 4 Distributions of MATTR with various window sizes (in brackets) for 1-grams across all subcorpus variants.

DOI: 10.22599/jesla.140

δ = 0.651, large), confounding the expectation that more expert writers would use a more diverse lexicon on average.

Interestingly, smaller window sizes show flatter curves towards increasing order. While this could be interpreted as a sign of decreased sensitivity to order, it should be remembered that our algorithm for artificially introducing order does so at a very specific scale, so this is more safely interpreted as an artefact.

5.1.2. Information-theoretic measures

Figure 5 shows the distribution for AIC and IFC. Both are clearly sensitive to increased order in the form of repetition, but respond to the exponential decrease of chunk size more mildly than TTR-based measures (which decrease exponentially). AIC is clearly insensitive to increased

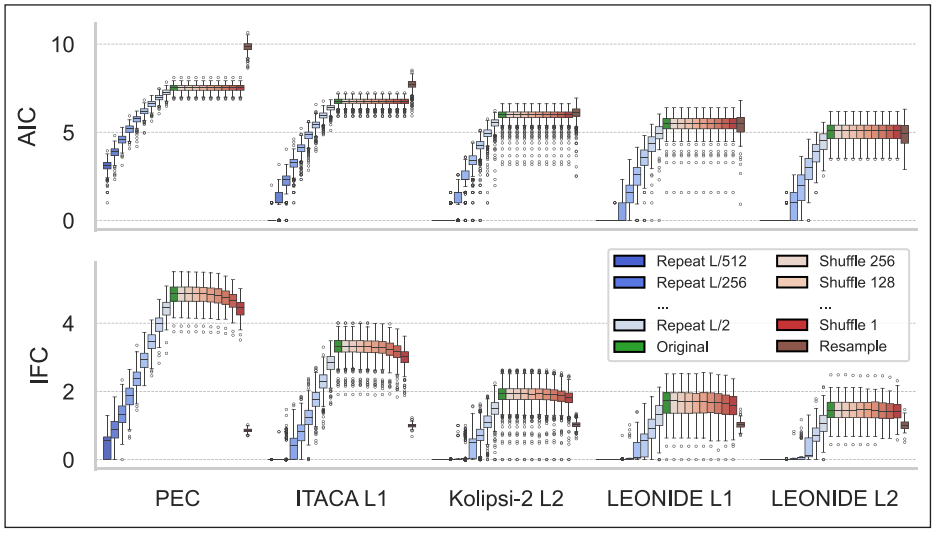

Figure 5 Information-theoretic measure distributions for 1-grams across all subcorpus variants.

disorder in the form of shuffling, as shown in the perfectly flat graph. This is expected, as it follows the bag-of-words model and is therefore by definition insensitive to the order of tokens.

IFC is sensitive to increased disorder in the form of shuffling, especially in subcorpora with longer text lengths. For the difference to be noticeable, the shuffled chunk needs to approach the size of a 4-gram for PEC and ITACA and of a unigram for Kolipsi-2. IFC is the only measure we test that is computed on adjacent token pairs and is thus expected to be sensitive to the order of tokens.

AIC and IFC are clearly sensitive to increased disorder in the form of uniform resampling. The uniformly resampled texts have tokens with probabilities that are almost all identical: in that case the value of AIC approaches $-log_2 N$ (thus the decreasing trend in magnitudes towards subcorpora with shorter texts) while the value of IFC approaches 0 (thus the seemingly identical small values across all subcorpora).

On the order spectrum, AIC exhibits a Kolmogorov-like behavior, while IFC is markedly Gell-Mann.

On the expertise spectrum, both AIC and IFC produce higher values for expert corpora, lower values for learner corpora and a clear difference between L1 corpora compared to L2 counterparts, as well as between older and younger writers.

Since IFC is dependent on *N*, it is possible for these results to change when controlling for text length. Figure 6 therefore displays results for MAIFC for a wide range of window sizes.

All caveats about window size choice discussed for MATTR also hold for MAIFC. Differences between subcorpora are smaller in magnitude, as expected. Furthermore, for very small window sizes (64, 128), Kolipsi-2 texts seem to have lower MAIFC values than for their younger L1 and L2 counterparts. The difference is however very small in magnitude (MATTR [window size 128]: Kolipsi-2 < LEONIDE L1, ***δ*** = 0.210, small; Kolipsi-2 < LEONIDE L2, ***δ*** = 0.196, small; MATTR [window size 64]: Kolipsi-2 < LEONIDE L1, ***δ*** = 0.217, small; Kolipsi-2 < LEONIDE L2, ***δ*** = 0.182, small). Additionally, it is interesting to note that while MATTR (window size 64) flattens the differences between original corpora visible in TTR, MAIFC (window size 64) still shows the

DOI: 10.22599/jesla.140

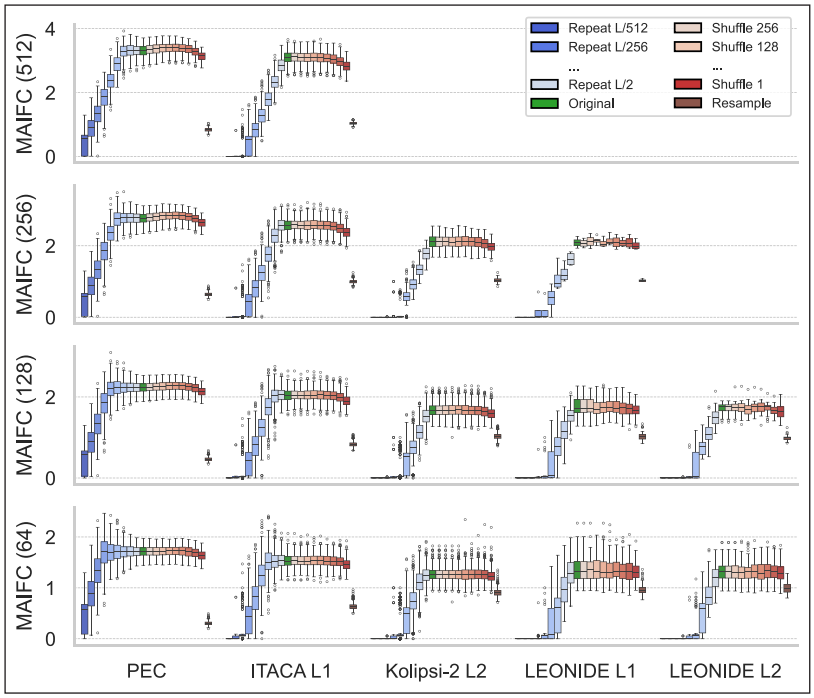

Figure 6 Distributions of MAIFC with different window sizes (in brackets) for 1-grams across all subcorpus variants.

expected differences between PEC and ITACA L1, as well as the other learner corpora, despite the very small window size.

5.2. MEASURES ON 2-GRAMS

We will now review the results for measures computed on 2-grams, focusing on the differences with the computations performed on 1-grams.

5.2.1. TTR-based measures

Figure 7 shows the distribution for TTR, RTTR and MATTR (window size 64). Unlike their 1-gram counterparts, TTR, RTTR and MATTR are all clearly sensitive to increased disorder, with larger effect sizes as the shuffled chunk decreases in size. This is the natural consequence

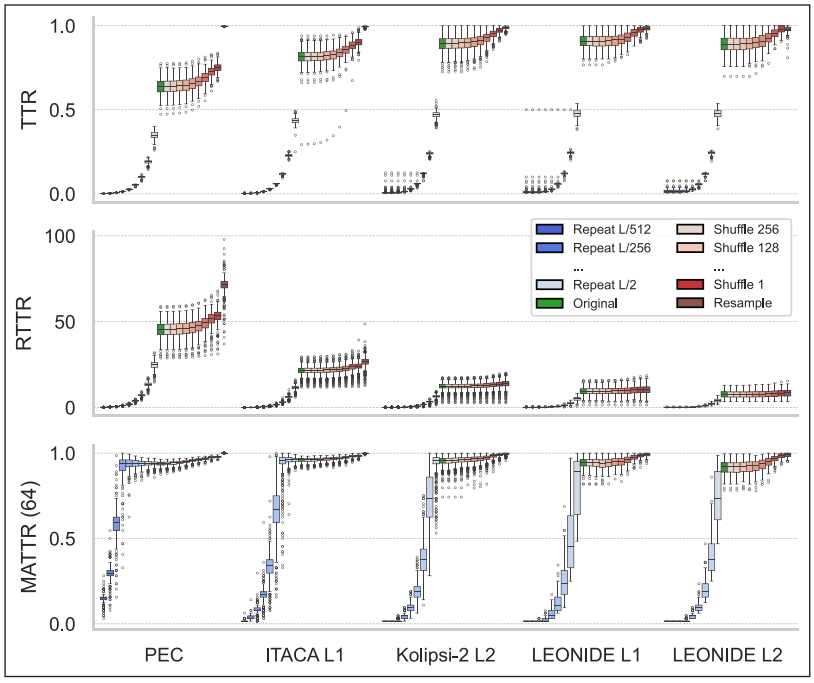

Figure 7 TTR-based measure distributions for 2-grams across all subcorpus variants.

of computing 2-grams: The bag-of-words model is no longer used and measures are now sensitive to shuffling. Furthermore, the values are closer to the maximum due to the fact that the 2-gram lexicon is much bigger than the 1-gram lexicon.

DOI: 10.22599/jesla.140

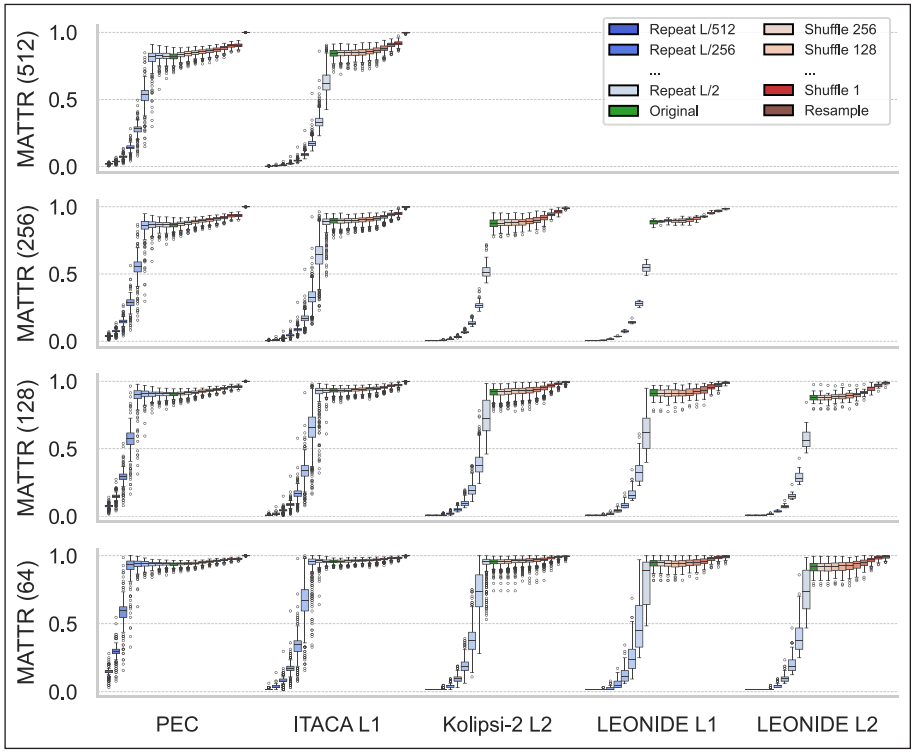

Figure 8 Distributions of MATTR with various window sizes (in brackets) for 2-grams across all subcorpus variants.

On the spectrum of order, TTR, RTTR and MATTR all exhibit markedly Kolmogorov behavior. On the expertise spectrum, the overall behavior of TTR and RTTR is not noticeably different from their 1-gram counterparts, while MATTR shows the anomalous effect of PEC texts having lower values than ITACA for all window sizes (see Figure 8).

The effect is barely noticeable from the graphs and requires further investigation using effect sizes. These are: MATTR (window size 512): ITACA L1 > PEC, $\delta = 0.430$, medium; MATTR (window size 256): $\delta = 0.618$, large; MATTR (window size 128): $\delta = 0.721$, large; MATTR (window size 64): $\delta = 0.788$, large. For 2-grams, PEC also shows lower MATTR values when compared to Kolipsi-2 and LEONIDE L1 texts, with effect sizes ranging from small to medium.

5.2.2. Information-theoretic measures

Figure 9 shows the distribution for AIC and IFC computed on 2-grams. For the same reasons as TTR-based measures, AIC also becomes sensitive to shuffling when computed on 2-grams. The effect is more visible for subcorpora with longer average text lengths (i.e., PEC, ITACA, Kolipsi-2), while for shorter text lengths, as in the case of LEONIDE data, shuffling effects are negligible for all shuffled chunk sizes. Thus, on the order spectrum, AIC exhibits a clear Kolmogorov behavior only when text lengths are longer on average. In comparison, IFC exhibits markedly Gell-Mann

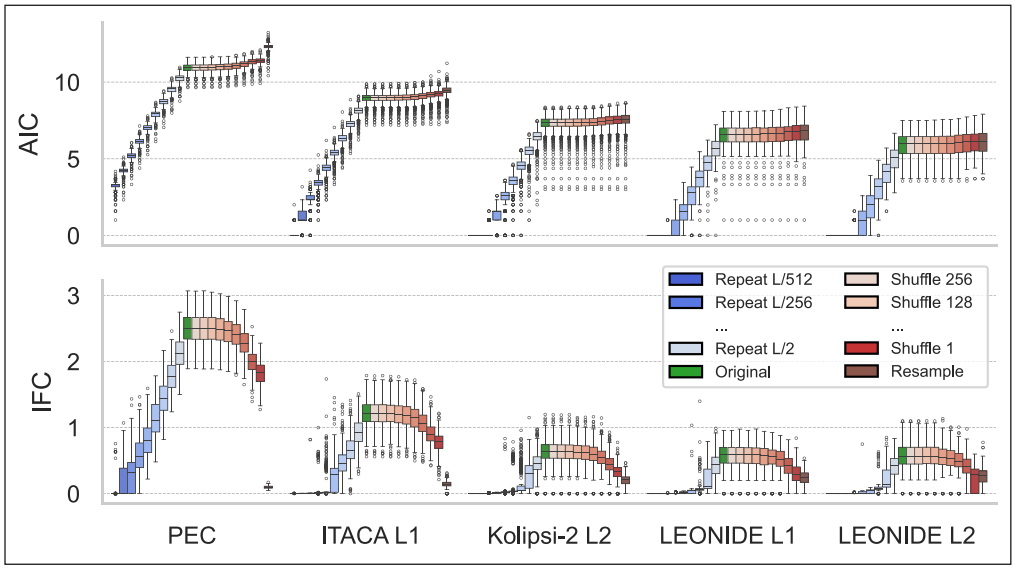

Figure 9 Information-theoretic measure distributions for 2-grams across all subcorpus variants.

DOI: 10.22599/jesla.140

behavior across all subcorpora, with a noticeable decrease in values starting at shuffled 8-grams in the case of PEC, ITACA and Kolipsi-2 and at 4-grams for LEONIDE L1 and L2 data. On the expertise spectrum, their overall behavior is not qualitatively different from their 1-gram counterparts.

Figure 10 shows the distribution of MAIFC for a wide range of window sizes. On the expertise spectrum, the overall behavior for small window sizes (64, 128) appears much flatter than the 1-gram counterpart, accentuating the counterintuitive differences between L1 and L2 corpora and across literacy levels discussed for unigrams (MAIFC [window size 128]: Kolipsi-2 < LEONIDE L1, $\delta = 0.144$, negligible; Kolipsi-2 < LEONIDE L2, $\delta = 0.509$, large; MAIFC [window size

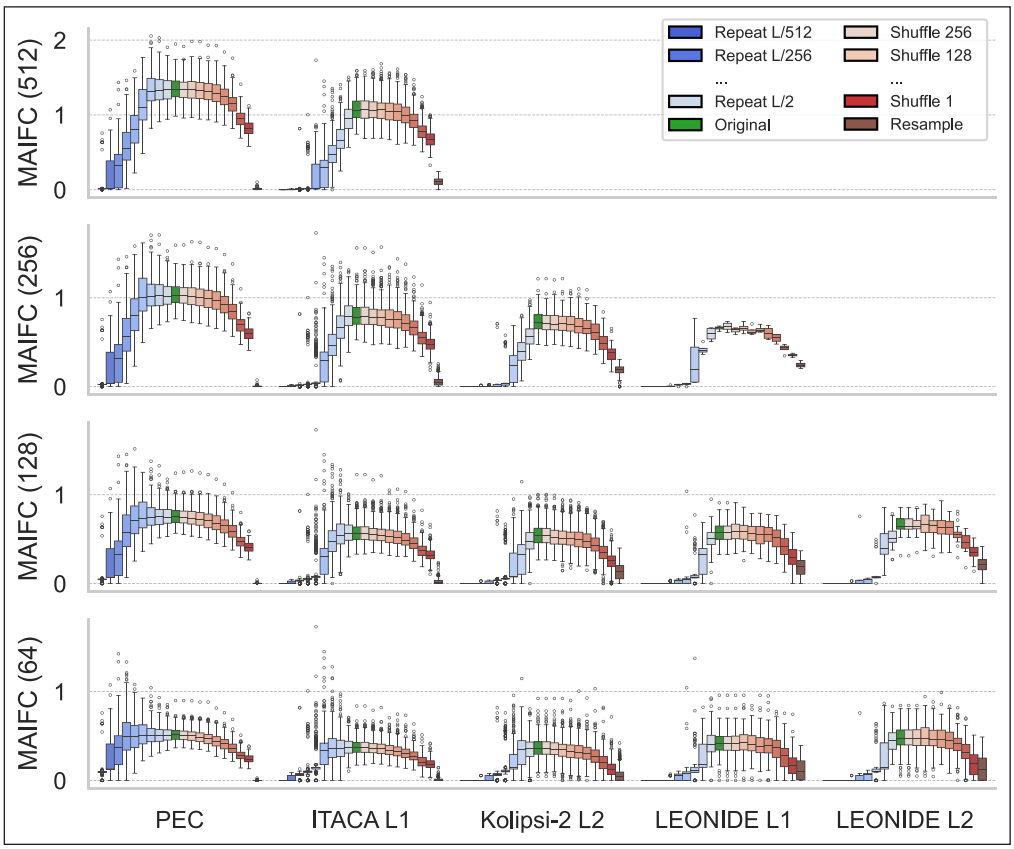

Figure 10 Distributions of MAIFC with different window sizes (in brackets) for 2-grams across all subcorpus variants.

64]: Kolipsi-2 < LEONIDE L1, $\delta = 0.246$, small; Kolipsi-2 < LEONIDE L2, $\delta = 0.517$, large). Across window sizes, PEC texts remain above the complexity levels displayed by learner corpora. All other considerations made for their 1-grams counterparts on the order spectrum still hold, apart from for one major difference: MAIFC now displays a very pronounced Gell-Mann behavior.

6. DISCUSSION

The simulation offered us a way to test which measures would be able to capture

- the expected variation across expertise in writing;
- the inter-relationships between constituent elements (and not only the number and variety) (Rescher, 1998);
- the notion of Gell-Mann complexity.

TTR, whether computed on 1-grams or on 2-grams, clearly showed a behavior that is not coherent with the assumption that lexicon size increases with expertise. This is easily explained by the fact that it ignores the fact that lexicon size grows in a non-linear way with respect to the text length, that is, Herdan-Heaps' empirical law is ignored.

MATTR, whether computed on 1-grams or on 2-grams, shows a decrease that is smaller in magnitude towards lower expertise. Furthermore, we observed a slightly anomalous effect: ITACA L1 texts consistently show more diversity than PEC data across window sizes and n-grams. We suggest this is because a given window size amounts to a different proportion

DOI: 10.22599/jesla.140

of the text depending on the length of the text itself, and thus captures different proportions of linguistically realized functions. For instance, a 512-token window in a typical ITACA L1 text, which is 600 words long, covers almost the whole text, which includes various functional components (e.g., setting the topic, taking a stance on it, presenting arguments and counter-arguments, concluding) that are typically expressed through different linguistic strategies in the text. In contrast, the same window in a 6000-word PEC text might only capture the introduction or a single paragraph of the body, which address fewer functional needs than a complete text. This thus yields lower diversity in the statistically averaged measures.

While MATTR shows blurred (1-grams) or counterintuitive (2-grams) distinctions between PEC and ITACA L1, MAIFC consistently distinguishes them in a predictable manner, even with small window sizes. Yet MAIFC behaves unexpectedly when comparing learner corpora: While other measures generally align with the intuitive expectation that texts from younger learners are less complex, Kolipsi-2 shows a slightly lower MAIFC than the LEONIDE counterparts. These differences are mostly small to negligible, except in one case: the comparison with LEONIDE L2 on 2-grams. While this small anomaly warrants further investigation, we note that it emerges more strongly on MAIFC computed on 2-grams which, unlike all other measures, does not lend itself to a straightforward linguistic interpretation.

Turning to how the measures capture inter-relationships between elements, we showed that TTR-based measures calculated on 1-grams cannot distinguish between original and shuffled variants as they are, by definition, a bag-of-words. Yet, when calculated on 2-grams, they reveal differences between original texts and shuffled variants, as bigrams are themselves inter-related units. Phraseological measures are thus an improvement over lexical measures because they better capture the complexity of word combinations arising from "the constraints on [a word's] co-occurrence with other words" (Paquot, 2019, p. 123).

Nonetheless, TTR-based measures and AIC show Kolmogorovian behavior on these interrelated units, meaning that these measures increase only in the case of a more disordered textual configuration. This invalidates the assumption or the expectation that diversity simply increases, with no upper limit, with proficiency, development or expertise. Instead, it explains the plateau often reported for phraseological diversity at higher proficiency levels or later developmental stages. This pattern likely reflects shifting writing conventions and training practices across different skill levels. Early language learners are often encouraged to avoid repetitions and showcase diverse vocabulary, while advanced writers, particularly in academic writing, are taught to prioritize clarity and coherence through consistent terminology. Repetition in this context reduces cognitive load, helping readers focus on meaning rather than decoding unnecessary lexical variation.

The notion of Gell-Mann complexity appears to be captured only by IFC (and MAIFC), both on 1-grams and 2-grams. Values clearly decrease towards both order and disorder, peaking in the middle range, where complex structures manifest – in other words, where the tension between conventionality and creativity is balanced.

A low Gell-Mann complexity can result from either extreme: excessive order or disorder. However, this ambiguity can be resolved by pairing it with a Kolmogorov measure, which increases monotonically with disorder. Together, Gell-Mann and Kolmogorov measures enhance each other's interpretability. Perceptual and qualitative assessments of a text's 'difficulty' can also enrich the interpretability of these quantities. We leave this endeavor for future research.

7. CONCLUSION

In this contribution, we first examined phraseological complexity across research literature, identifying how the construct moves from conceptualization to operationalization and cataloguing common measurement strategies. We then tested the most widely used measures for computing the diversity component of phraseological complexity, leveraging the controlled setting afforded by simulations. Working from the hypothesis that writers neither become more orderly nor more disorderly as they develop cognitively or gain expertise in a specific genre or advance in proficiency, we found an intuitive way to artificially recreate order and disorder in their productions and test which of the widely used and innovative measures would be able to catch this trend. Our findings show that IFC is the only measure that peaks for

DOI: 10.22599/jesla.140

original texts, when writers supposedly strike a balance between creativity and conventionality. We recommend pairing this Gell-Mann measure with Kolmogorov measures to afford even more explainability of low scores.

To frame the scope of our findings more clearly, we would like to point to some limitations and future directions worth exploring. First, the datasets we chose for the analysis provide only cross-sectional and pseudo-longitudinal insights, which cannot be generalized to proficiency levels. Testing these new measures against proficiency would require datasets with appropriate metadata. With regards to the newly proposed measures, information-theoretic measures like IFC still lack clear linguistic interpretation, since they do not stem from nor entail linguistic theories. While these measures facilitate cross-disciplinary communication since they use minimal assumptions about the code to be quantified, they do not as yet map to intuitive linguistic features. Significant theoretical and empirical work remains to establish these connections.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their brilliant suggestions, which improved the robustness of the reported results and their interpretation. We also thank Stefania Spina for access to the Perugia Corpus academic prose subsection and Jennifer-Carmen Frey for her valuable feedback on a previous version of this manuscript. Any remaining errors are our own. Publication in Open Access was supported by the Eurac Research Open Access Fund.

AUTHOR AFFILIATIONS

Paolo Brasolin orcid.org/0000-0003-2471-7797
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione, Università Cattolica del Sacro Cuore, Milano, Italy

Arianna Bienati orcid.org/0000-0002-5732-3957
Dipartimento di Educazione e Scienze Umane, Università degli studi di Modena e Reggio Emilia, Reggio Emilia, Italy; Institute for Applied Linguistics, Eurac Research, Bolzano, Italy

REFERENCES

Ackermann, K., & **Chen, Y.-H.** (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Alzahrani, A. (2024). Utility of Kolmogorov complexity measures: Analysis of L2 groups and L1 backgrounds. *PLOS ONE*, *19*(4), Article e0301806. https://doi.org/10.1371/journal.pone.0301806

Bates, J. E. (2020). *Measuring complexity using information fluctuation: A tutorial*. ResearchGate. https://www.researchgate.net/publication/340284677_Measuring_complexity_using_information_fluctuation_a_tutorial

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

Bienati, A., Frey, J.-C., Zanasi, L., Stemle, W. E., Palmero Aprosio, A., Brasolin, P., & **Vettori, C.** (forth.). *ITACA – A corpus of Italian argumentative student essays for the evaluation of coherence*. CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy, Cagliari.

Biondi, G., Franzoni, V., Milani, A., & **Santucci, V.** (2023). Classification of Text Writing Proficiency of L2 Learners. In O. Gervasi, B. Murgante, A. M. A. C. Rocha, C. Garau, F. Scorza, Y. Karaca, & C. M. Torre (Eds.), *Computational Science and Its Applications – ICCSA 2023 Workshops* (pp. 15–28). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37105-9_2

Bottini, R., & **Le Foll, E.** (2024). The more proficient the learners, the less sophisticated their L2 vocabulary?: The curious effect of the reference corpus on mean-frequency measures of lexical sophistication. *International Journal of Learner Corpus Research*, *11*(1), 47–78. https://doi.org/10.1075/ijlcr.23029.bot

Brezina, V., & **Pallotti, G.** (2019). Morphological complexity in written L2 texts. *Second Language Research*, *35*(1), 99–119. https://doi.org/10.1177/0267658316643125

Bulté, B., & **Housen, A.** (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & **Khalil, H.** (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different— the "Big Picture" review family. *Systematic Reviews*, *12*(1), Article 45. https://doi.org/10.1186/s13643-023-02178-5

Clercq, B. D., & **Housen, A.** (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research*, *35*(1), 71–97. https://doi.org/10.1177/0267658316674506

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Corral, Á., & **Font-Clos, F.** (2017). Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Physical Review E*, *96*(2), Article 022318. https://doi.org/10.1103/PhysRevE.96.022318

Dahl, Ö. (2009). Testing the assumption of complexity invariance: The case of Elfdalian and Swedish. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language Complexity as an Evolving Variable* (pp. 50–63). Oxford University Press. https://doi.org/10.1093/oso/9780199545216.003.0004

Ehret, K., & **Szmrecsanyi, B.** (2016). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research*, *35*(1), 23–45. https://doi.org/10.1177/0267658316669559

Esfandiari, R., & **Ahmadi, M.** (2023). Phraseological Complexity and Academic Writing Proficiency in Abstracts Authored by Student and Expert Writers. *English Teaching & Learning*, *47*(4), 429–448. https://doi.org/10.1007/s42321-022-00118-5

Forti, L. (2020). L2 phraseology research at the interface between learner corpus research and psycholinguistics. *Rivista Di Psicolinguistica Applicata*, *20*(2), 19–33. https://doi.org/10.19272/202007702002

Futrell, R., & **Hahn, M.** (2022). Information Theory as a Bridge Between Language Function and Language Form. *Frontiers in Communication*, *7*, Article 657725. https://doi.org/10.3389/fcomm.2022.657725

Gell-Mann, M. (1994). *The quark and the jaguar: adventures in the simple and the complex*. Little Brown. https://doi.org/10.1063/1.2808634

Gell-Mann, M., & **Lloyd, S.** (1996). Information measures, effective complexity, and total information. *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

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & **Levy, R.** (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Glaznieks, A., Frey, J.-C., Abel, A., Nicolas, L., & **Vettori, C.** (2023). The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German. *Italian Journal of Computational Linguistics*, *9*(2), Article 2. https://doi.org/10.4000/ijcol.1210

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

Goldberg, A. E. (2013). Constructionist Approaches. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 14–31). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0002

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: Essai de méthodologie*. Presses universitaires de France.

Guzzi, E., & **Alonso Ramos, M.** (2023). Sophistication and Diversity as Lexical Complexity Measures to Identify the Collocational Profile of Spanish Academic Texts. *Revista Signos*, *56*(112), 282–305. https://doi.org/10.4067/S0718-09342023000200282

Hu, R., Wu, J., & **Lu, X.** (2022). Word-Combination-Based Measures of Phraseological Diversity, Sophistication, and Complexity and Their Relationship to Second Language Chinese Proficiency and Writing Quality. *Language Learning*, *72*(4), 1128–1169. https://doi.org/10.1111/lang.12511

Jiang, J., Bi, P., Xie, N., & **Liu, H.** (2023). Phraseological complexity and low- and intermediate-level L2 learners' writing quality. *International Review of Applied Linguistics in Language Teaching*, *61*(3), 765–790. https://doi.org/10.1515/iral-2019-0147

Kim, M., & **Crossley, S. A.** (2023). Lexical and phraseological differences between second language written and spoken opinion responses. *Frontiers in Psychology*, *14*, Article 1068685. https://doi.org/10.3389/fpsyg.2023.1068685

Kolmogorov, A. N. (1963). On Tables of Random Numbers. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *25*(4), 369–376.

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93. https://doi.org/10.1515/lingvan-2021-0112

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford University Press.

Mangiafico, S. S. (2016). *Summary and Analysis of Extension Program Evaluation in R*. Rutgers Cooperative Extension. https://rcompanion.org/documents/RHandbookProgramEvaluation.pdf

Massip-Bonet, À., Bel-Enguix, G., & **Bastardas-Boada, A.** (Eds.). (2019). *Complexity Applications in Language and Communication Sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-030-04598-2

Mastrantonio, D. (2021). L'italiano scritto accademico: Problemi descrittivi e proposte didattiche. *Italiano LinguaDue*, *13*(1), 348–368.

Milani, A., Franzoni, V., & **Biondi, G.** (2021). Parsing tools for Italian phraseological units. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, & C. M. Torre (Eds.), *Computational science and its applications – ICCSA 2021, Part VII* (pp. 427–435). https://doi.org/10.1007/978-3-030-87007-2_30

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117–134. https://doi.org/10.1177/0267658314536435

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Paquot, M., Naets, H., & **Gries, S. T.** (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb plus object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 122–147). https://doi.org/10.1017/9781108674577.007

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1), 219–242. https://doi.org/10.1016/S0024-3841(98)00035-7

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & **Manning, C. D.** (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Rescher, N. (1998). *Complexity: A Philosophical Overview*. Routledge. https://doi.org/10.4324/9780429336591

Romano, J., Kromrey, J. D., Coraggio, J., & **Skowronek, J.** (2006). *Appropriate Statistics for Ordinal Level Data: Should We Really Be Using t-test and Cohen's d for Evaluating Group Differences on the NSSE and other Surveys?* Annual meeting of the Florida Association of Institutional Research.

Rubin, R. (2021). Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in L2 Dutch. *International Journal of Learner Corpus Research*, *7*(1), 131–162. https://doi.org/10.1075/ijlcr.20005.rub

Rubin, R., Housen, A., & **Paquot, M.** (2021). Phraseological Complexity as an Index of L2 Dutch Writing Proficiency: A Partial Replication Study. In S. Granger (Ed.), *Perspectives on the L2 Phrasicon: The view from learner corpora* (pp. 101–125). Multilingual Matters. https://doi.org/10.21832/9781788924863-006

Serra-Peralta, M., Serrà, J., & **Corral, Á.** (2021). Heaps' law and vocabulary richness in the history of classical music harmony. *EPJ Data Science*, *10*(1), Article 1. https://doi.org/10.1140/epjds/s13688-021-00293-8

Spina, S. (2014). Il Perugia Corpus: Una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-It 2014*, *1*, 354–359. https://hdl.handle.net/20.500.12071/2321

Torchiano, M. (2020). *effsize: Efficient Effect Size Computation* (Version 0.8.1) [Computer software]. https://cran.r-project.org/web/packages/effsize/index.html

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

Vandeweerd, N., Housen, A., & **Paquot, M.** (2023). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*, *45*(4), 1–25. https://doi.org/10.1017/s0272263122000389

Wackerbauer, R., Witt, A., Atmanspacher, H., Kurths, J., & **Scheingraber, H.** (1994). A comparative classification of complexity measures. *Chaos, Solitons & Fractals*, *4*(1), 133–173. https://doi.org/10.1016/0960-0779(94)90023-X

Wang, G., Wang, H., & **Wang, L.** (2022). Kolmogorov complexity metrics in assessing L2 proficiency: An information-theoretic approach. *Frontiers in Psychology*, *13*, Article 1024147. https://doi.org/10.3389/fpsyg.2022.1024147

Yin, X., Cao, G., Wang, L., & **Xu, J.** (2024). Automatic Readability Assessment Based on Phraseological Complexity. In J. Gan, Y. Pan, J. Zhou, D. Liu, X. Song, & Z. Lu (Eds.), *Computer Science and Educational Informatization* (pp. 3–11). Springer Nature. https://doi.org/10.1007/978-981-99-9492-2_1

Zhang, Y., & **Ouyang, J.** (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, *56*, Article 100727. https://doi.org/10.1016/j.asw.2023.100727

TO CITE THIS ARTICLE:
Brasolin, P., & Bienati, A. (2025). Phraseology meets information theory: Going beyond the *bag-of-words* approach in complexity measures. *Journal of the European Second Language Association*, 9(1), 103–123. DOI: https://doi.org/10.22599/jesla.140

Submitted: 30 September 2024
Accepted: 11 April 2025
Published: 04 July 2025

WHITE ROSE UNIVERSITY PRESS
Universities of Leeds, Sheffield & York