

Phraseological complexity measures in learner Italian. Integrating eye tracking, computational and learner corpus methods to develop second language pedagogical resources.



PHRAME

Phraseological complexity
measures in learner Italian

CLOSING CONFERENCE
27-28 November 2023

Book of Abstracts

Table of Contents

Scientific and organising committee.....	3
Keynote presentations	4
<i>Learner Corpora and Phraseology</i> , Philip Durrant.....	5
<i>Defining, measuring and interpreting complexity</i> , Gabriele Pallotti	6
<i>Phrasal processing: Past, present, and future</i> , Anna Siyanova-Chanturia	7
Paper presentations	8
<i>Exploring phrasal verbs of action and motion in learner and native speaker narrative writing</i> , Katherine Ackerley and Erik Castello.....	9
<i>Reading versus listening to authentic materials in L2 Italian: What leads to greater vocabulary learning gains?</i> Mahnaz Aliyar and Anna Siyanova-Chanturia	10
<i>Exploring Phraseological Patterns in English Non-Finite Clauses: A Corpus-Based Study of Business English</i> , Olfa Ben Amor	11
<i>Phraseological Patterns in Learner Academic English: Insights from Corpus-Driven Approaches</i> , Sibel Aybek and Cem Can	12
<i>Between order and disorder: an ecological view on lexical diversity measures</i> , Arianna Bienati and Paolo Brasolin	13
<i>The ‘growth’ of academic phrases: a contrastive corpus-based study on phraseological complexity in Romanian learner English</i> , Madalina Chitez, Andreea Dinca, Ana-Maria Bucur, and Kristian Miok.....	15
<i>Discovering the potential of automated phraseological interference error detection: Transformer-based approach</i> , Darya Kharlamova	17
<i>The effect of the reference corpus on lexical and phraseological sophistication measures: Reflecting on their reliability and validity in L2 English</i> , Elen Le Foll and Raffaella Bottini	19
<i>Phraseological complexity measures and the assessment of L2 writing</i> , Agnieszka Leńko-Szymańska, Piotr Peżnik, and Michał Adamczyk.....	20
<i>The figurative and phraseological component of Italian academic register</i> , Davide Mastrantonio	21
<i>Torn between L1 and L2 patterns? Collocationality levels in L2 English production by speakers of L1 Italian</i> , Maja Miličević Petrović, Adriano Ferraresi, and Silvia Bernardini	23
<i>Collocation processing in writing and translation between Chinese and English:</i>	25
<i>A corpus-based and keylogging analysis</i> , Qiuqing Qin.....	25
<i>Bottom-up meets top-down: Investigating the construct validity of phraseological complexity measures through a cross-methodological comparison</i> , Nathan Vandeweerd, Fanny Forsberg Lundell, and Klara Arvidsson	27

Eurac Research, Bolzano

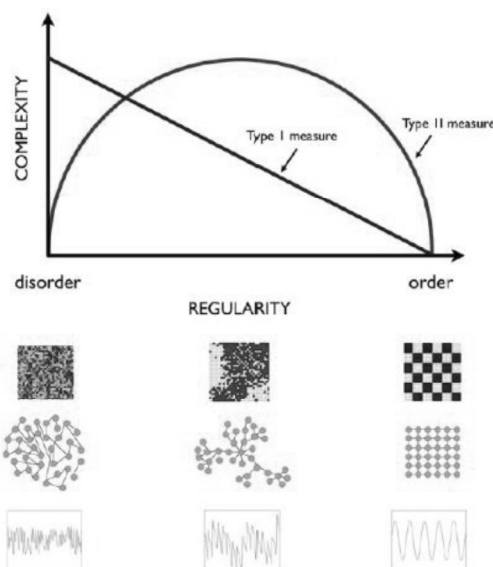
Starting from the influential study by Paquot (2019), many scholars have conceptualised phraseological complexity as composed by two intertwined components: phraseological diversity and sophistication (e.g., Vandeweerd et al., 2022). To compute phraseological diversity, lexical diversity measures such as root type-token ratio have been used. Although there have been significant contributions to the debate on the validity of such indices (Kyle et al., 2021; McCarthy & Jarvis, 2010), a thorough comparison with theoretical observations from other fields of research would greatly enhance our understanding of the specific type of complexity that these measures truly capture.

Typological literature has distinguished two types of complexity (Dahl, 2009): Kolmogorov complexity refers to the length of the shortest description needed to represent a string of symbols. Gell-Mann effective complexity is understood as the length of description required to specify the set of regularities present in a string. According to Kolmogorov complexity, a string that lacks any regular patterns or structure would require a very long thus complex description, as each element of the string would need to be individually accounted for. From the perspective of Gell-Mann effective complexity, instead, a string without any regularities would have no complexity, since it contains no structured patterns.

In ecology (Parrott, 2010), a clear-cut distinction is made between measures that operationalise the Kolmogorov complexity (“Type 1 measures”) vs. the Gell-Mann effective complexity (“Type 2 measures”). Type 1 measures favor random sequences and can be visualised as linear functions on the continuum between order and disorder; type 2 measures, instead, can be visualised as a convex function that reaches its peak when a system strikes a balance between rules and exceptions (Fig. 1). Intuitively, the latter might be the kind of complexity we want to be able to measure in a text.

This contribution will therefore try to answer the following research question: which type of complexity (Kolmogorov vs. Gell-Mann) is measured by the most commonly used lexical diversity indices?

In addressing this question, we will conduct a simulation to evaluate lexical diversity in Italian texts employing TTR-based metrics along with entropy-based ones (as found in Garner, 2020), including Type 2 measures, such as *fluctuation complexity*, as described in Parrot (2010). Each measure will be calculated varying the text length by truncation to study the length dependency of its behavior. Additionally, each measure will be calculated on texts which have different expected behaviors in terms of Kolmogorov and Gell-Mann complexity: real texts –



expert-authored (e.g., Repubblica corpus) and non-expert-authored (e.g., LEONIDE, ITACA)¹ – are expected to display a certain degree of regularity, thus being representative of a medium to high Gell-Mann complexity; synthetic texts sampling the real texts using either the original word distributions or a uniform distribution (i.e., every word has an equal probability of appearing), instead, do not show this regularity, thus representing a medium to high Kolmogorov complexity, but a low Gell-Mann complexity. We will then be able to distinguish between type 1 and type 2 measures by identifying the set of texts (real or synthetic) for which the various lexical diversity measures yield the highest values. Results will contribute to the assessment of metrics’ validity for different conceptualisations of complexity.

References

- 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.
- 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.
- Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, 18(2), 154–170.
- McCarthy, P. M., & Jarvis, S. (2010). MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392.
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35, 121–145.
- Parrott, L. (2010). Measuring ecological complexity. *Ecological Indicators*, 10(6), 1069–1076.
- 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.

¹ Size and scale of cited corpora:

Corpus	Size (in n. of tokens)	N. of documents
Repubblica corpus (from: http://sslmit.unibo.it/repubblica)	380,823,725	572,515
LEONIDE_IT (from: http://hdl.handle.net/20.500.12124/25)	93,000	844
ITACA (not yet published)	382,964	635