

Disentangling the Construct of Phraseological Sophistication: a Validation Study with Human Judgments

Work-in-progress Report
LxGr26

Gabriela Vaughan and Magali Paquot
July, 2026

Overview

01 Background

02 RQs

03 Methods

04 Analysis

05 Preliminary results

06 Take-home message

Background

Linguistic Complexity is

- Major variable in SLA research
- Used to:
 - describe L2 performance
 - assess L2 proficiency
 - track L2 development

(Housen & Kuiken, 2009; Paquot, 2019)

Problem

- Focus on lexical and syntactic complexity measures
- There remains a need for **measures of complexity** that capture language at the **phrasal level** (Paquot, 2019).

Why should we care about phraseology (and its measures)?

Phraseology plays an essential role in **language proficiency and development** (Wray, 2002).

Research shows that more proficient learners use a wider range of strongly associated collocations (Bestgen & Granger, 2018; Paquot, 2019).

Phraseological Complexity = range (diversity) +
sophistication of phraseological units

(Paquot, 2019)

A Focus on Sophistication

- **Why Phraseological Complexity?**
 - Phraseological Complexity has emerged as a valuable construct that provides insights into the description, assessment and tracking of learner language
 - Multiple studies have established the role of PC across a variety of L2 groups
 - **Phraseological Sophistication** is a particularly promising measure:
 - Automatic measures of PS can distinguish between proficiency (CEFR) levels (predictive validity), highlighting the potential of phraseological complexity as a measure of L2 development.

The dimensions of phraseological sophistication

- Register specificity: Academic vs. Non-Academic Collocations
 - Association strength (MI): How strongly words co-occur
 - Frequency: Common vs. rare combinations
- (Paquot & Naets, 2025)

The theoretical construct assumes that collocations which are more sophisticated are:

register-specific
strongly associated
less frequent



Research Gap

Currently

- Phraseological complexity (including phraseological sophistication) is typically measured using **automatic corpus-based indices**.

Unclear

- Whether these measures reflect how **humans** actually **perceive** phraseological complexity (and sophistication).

Problem

- Construct validity and convergent validity in Applied linguistics: linguistic **constructs** are **rarely validated through human perception studies** (Jarvis, 2017).
- .



This study aims to empirically test whether reliable human judgments of sophistication are attainable at the phrasal level and determine which dimensions of phraseological sophistication human judges are sensitive to.

Research Questions

- **RQ1:** To what extent do L1 speakers of English share similar intuitions about the sophistication of word combinations?
- **RQ2:** To what extent do dimensions of phraseological sophistication influence human judgments of phraseological sophistication?

Method: Comparative Judgment

Comparative Judgment (CJ) originates in the field of Psychology (Thurstone, 1927). This is a novel approach to evaluating complex linguistic constructs (Thwaites & Paquot, 2024).

Judges compare two items:

“Which word combination is more sophisticated?”

text A or *text B*



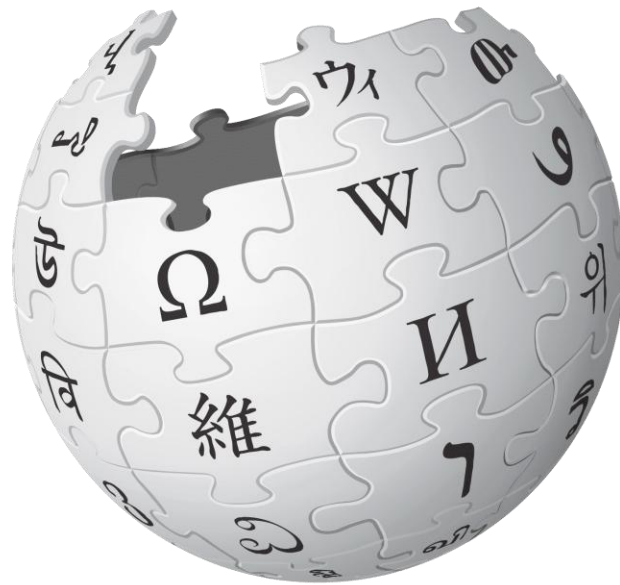
Comparative Judgment

Pros:

- ✓ Requires minimal training with clear instructions
- ✓ Reliability: it aggregates multiple judgments, reducing individual bias
- ✓ The holistic nature of the task better captures the multidimensional nature of linguistic constructs.
- ✓ Promotes construct validity as it captures a collective interpretation of a linguistic construct



Verb + direct object collocations



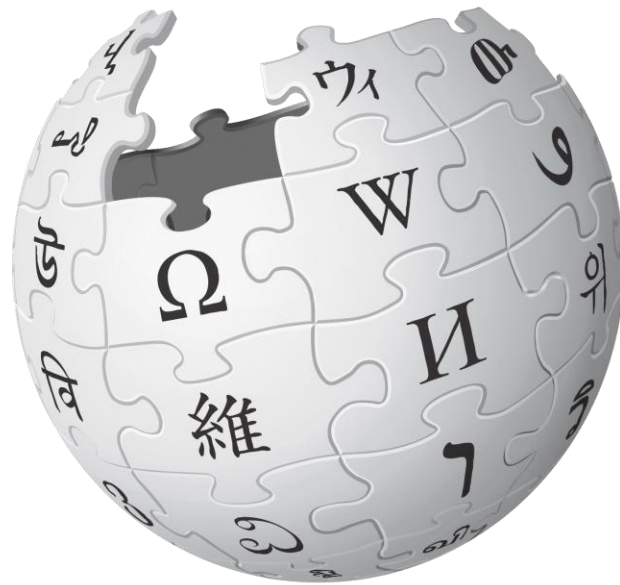
WIKIPEDIA
The Free Encyclopedia

Corpus

- WikiCorpus
- 6.8M texts

Extracted

- Verb + direct object collocations
- Organised using hierarchical word tiers (Burch & Egbert, 2023)



WIKIPEDIA
The Free Encyclopedia

Dataset

- **128 collocations**
- Balanced across dimensions of phraseological sophistication
 - Register specificity
 - Association strength
 - Frequency

Dimensions of phraseological sophistication

Dimensions	Operationalisation	+	-
Register Specificity	Academic vs. Non-Academic Collocation	<i>acquire knowledge</i> (ACL)	<i>find information</i> (Non-ACL)
Association Strength	How strongly words co-occur (PMI score)	<i>shed light</i> (PMI = 9.8)	<i>present case</i> (PMI = 3)
Frequency	Common vs. Rare Collocations	<i>provide insight</i> (occurs 6,803 times)	<i>corroborate finding</i> (occurs 91 times)

Study Design

Design

- **30 expert judges (L1 English)** make **pairwise comparisons** between two collocations and decide which better exemplifies phraseological sophistication
- 50 comparisons each
- 1,500 comparisons total
- Feedback form collecting qualitative comments about characteristics of collocations that assisted in their decision-making process



Judges' instructions

In this study, the term is used in a broad sense, reflecting how it is commonly applied in research on vocabulary and phraseology. In general, phraseological sophistication refers to differences in how certain word combinations are considered conventional or typical in language use. One way to think of phraseological sophistication is that some word combinations may appear more idiomatic, conventional, or natural in context than alternative combinations.

This broad framing is intended to provide a minimal common starting point for the task, rather than a strict definition. As there is no clear consensus on the definition of phraseological sophistication, we invite you to rely primarily on your intuition and professional experience when deciding which word combinations you consider more phraseologically sophisticated.



Judges compare two items:

< Left

2.4 secs

Which word combination is... | Undo

49

Right >

take possession

investigate
impact



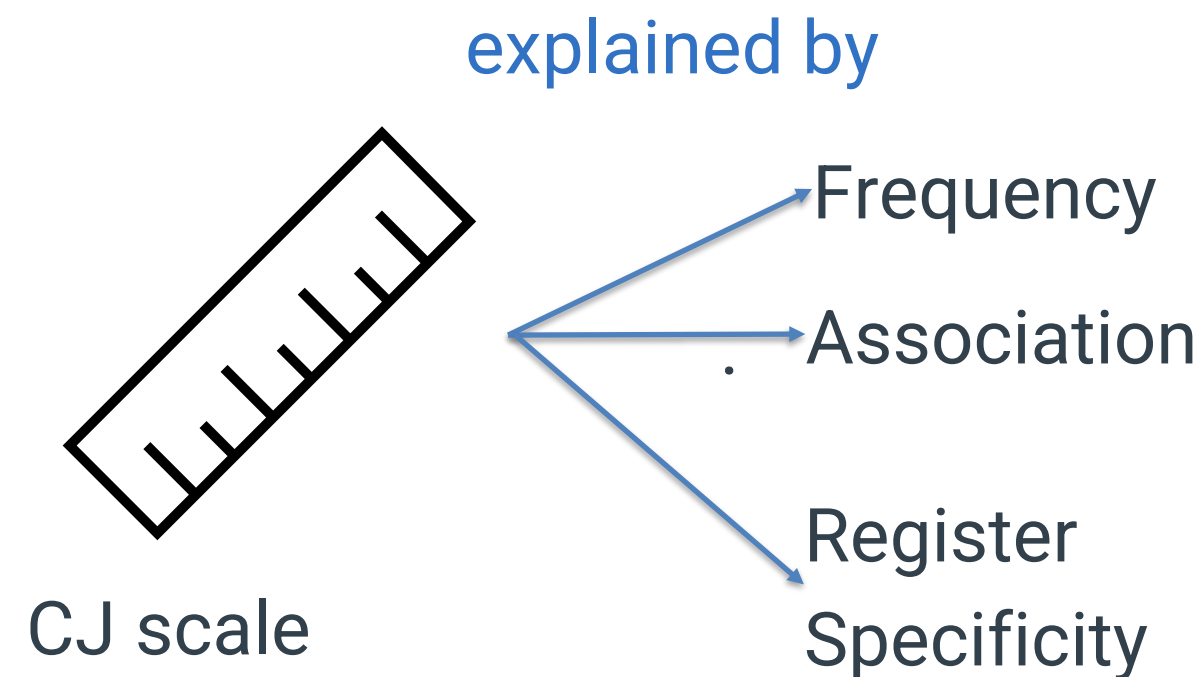
Output

- **Scale ranking** 128 collocations from **least sophisticated to most sophisticated**

least sophisticated collocation  most sophisticated collocation



Analysis



RQ1: Can phraseological sophistication be judged reliably?

- Assess inter-judge reliability using Scale Separation Reliability (SSR).

RQ2: What drives perceptions of phraseological sophistication?

- Generate a Comparative Judgment (CJ) scale ranking collocations from least to most sophisticated.
- Determine alignment with corpus-based measures
 - Calculate indices of frequency, association and register specificity
- Use regression modelling to determine the extent to which these dimensions predict the CJ scale.
- Thematic analysis of qualitative comments regarding the decision-making process

Preliminary Results

Data collection concluded on 28/06/2026

- 30 judges recruited, 28 completed the CJ task
- Current reliability: 0.85
 - Indicates a good level of consistency in their decisions

Sample of qualitative comments

What characteristics of the word combinations did you rely on when making decisions during the experiment?

“based on my subjective feelings of whether a word/phrase felt 'academic' or more 'every day’”

“Some decisions were based on single words and others on the full phrase”

“ academic appearance .”

“ Research related phrases.”

“Phrases suggesting more abstract concepts would seem to be more sophisticated, as would those where a collocation seems less literal or transparent in meaning.”

Take-home message

- Phraseological sophistication is a promising dimension of complexity, but its construct validity remains largely untested.
 - This research investigates whether human judges share a common understanding of phraseological sophistication.
 - By comparing human judgments with corpus-based measures, this study will provide evidence for the current operationalisation of phraseological sophistication.
-

Thank you for your time 😊

Selected References

- Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4), 324. <https://doi.org/10.2307/2334029>
- Crossley, S. (n.d.). *Crowd-Sourcing Human Ratings of Linguistic Production*.
- Fearnley, A. (2000). A comparability study in GCSE mathematics. A study based on the summer 1998 examination. Assessment and Qualifications Alliance (Northern Examinations and Assessment Board). Manchester: Joint Forum for the GCSE and GCE.
- Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, 34(4), 537–553. <https://doi.org/10.1177/0265532217710632>
- Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY, USA: John Wiley & Sons Inc.
- Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating. *Language Learning*, 72(3), 853–885. <https://doi.org/10.1111/lang.12498>
- Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281–300. <https://doi.org/10.1080/0969594X.2012.665354>
- Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 166–206. London: QCA.
- Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the Validity of Quantitative Measures in Applied Linguistics Research¹. *Language Learning*, 65(S1), 37–75. <https://doi.org/10.1111/lang.12112>
- Robitzsch, A. (2022). SIRT: Supplementary Item Response Theory Models. <https://CRAN.R-project.org/package=sirt>.
- Thwaites, P., & Paquot, M. (2024). Comparative judgement for advancing research in applied linguistics. *Research Methods in Applied Linguistics*, 3(3), 100142. <https://doi.org/10.1016/j.rmal.2024.100142>
- Thwaites, P., & Paquot, M. (2025). Testing crowdsourcing as a means of recruitment for the comparative judgement of L2 argumentative essays. *Journal of Second Language Writing*, 68, 101207. <https://doi.org/10.1016/j.jslw.2025.101207>
- 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), 787–811. <https://doi.org/10.1017/S0272263122000389>
- Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5), 541–562. <https://doi.org/10.1080/0969594X.2019.1602027>
- Zhang, X., & Lu, X. (2024). Testing the Relationship of Linguistic Complexity to Second Language Learners' Comparative Judgment on Text Difficulty. *Language Learning*, 74(3), 672–706. <https://doi.org/10.1111/lang.12633>
- Zhang, Y., & Ouyang, J. (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, 56, 100727. <https://doi.org/10.1016/j.asw.2023.100727>