Measuring Lexicogrammatical Complexity and Sophistication in Second Language English Production: Development and Validation of Argument Structure Construction-based Indices

by

Hakyung Sung

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Linguistics

Dissertation Committee:

Kristopher Kyle, Chair

Vsevolod M. Kapatsinski, Core Member

Don Daniels, Core Member

Thien Huu Nguyen, Institutional Representative

University of Oregon

Spring 2025

DISSERTATION ABSTRACT

Name: Hakyung Sung

Degree: Doctor of Philosophy in Linguistics

Title: Measuring lexicogrammatical complexity and sophistication in second language English production: Development and validation of argument structure construction-based indices

Grounded in a usage-based constructionist framework, this study conceptualizes language as a network of entrenched form-meaning pairings (i.e., constructions) that shape individual grammars through repeated exposure. While traditional measures of syntactic complexity capture some facets of L2 proficiency, they seldom treat constructions as the focus of analysis. Indices derived from argument structure constructions (ASCs), which map syntactic arguments onto semantic roles and thus encode fundamental human experiences, provide a complementary perspective. Nevertheless, even though studies consistently show that learners' ASC usage becomes more complex and sophisticated with increasing proficiency, scalable and systematic methods for extracting and analyzing ASC-based indices remain limited.

To address this gap, the present study introduces ASC analyzer, an open-source NLP tool that builds on a RoBERTa-based ASC tagger trained on a gold-standard treebank of L1 and L2 English. The analyzer automatically labels ASCs and computes a suite of ASC-based indices (i.e., diversity, proportion, frequency, and verb-construction strength of association) for large-scale corpus analyses.

Empirical validation in an L2 speaking-assessment task shows that these ASC-based indices yield nuanced insights into how learners at different oral-proficiency levels deploy constructions and verbs while completing the same task and possess solid predictive power for speaking scores. When combined with additional lexicogrammatical measures, they further boost the model's explanatory power. A parallel study of L2 writing corroborates these findings: adding ASC-based indices not only outperforms traditional syntactic-complexity metrics in isolation but also enhances models that already include syntactic and lexicogrammatical predictors. The results demonstrate that ASC-based analysis offers a valuable contribution to multivariate frameworks that seek to capture the complex interplay between grammatical form and lexical choice in L2 production.

This dissertation includes previously published and unpublished co-authored materials.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. Kris Kyle, whose mentorship has shaped both my academic and personal growth. Through countless conversations, hands-on guidance, and unwavering support, Kris has shown me that becoming a good researcher begins with becoming a good person. I've learned not only how to conduct research, but also how to lead a balanced, thoughtful life in academia. I cannot thank him enough for the many lessons I've come to understand simply by doing, with his encouragement behind me and his example ahead of me.

I also wish to thank my committee members: Dr. Volya Kapatsinski, Dr. Don Daniels, and Dr. Thien Nguyen. Volya and Don have supported me since my second qualifying paper, which laid the groundwork for this dissertation's work on ASC annotation. Volya's sharp questions and close reading of the manuscript challenged me to think more precisely and critically. Don's feedback on constructions, syntax, and semantics helped me rethink key concepts central to this study. Thien's courses and seminars in the computer science department deepened my understanding of natural language processing. Though I am still learning, much of the knowledge I now draw upon stems from his supportive instruction.

Heartfelt thanks also go to the many others who supported me along the way: my cohorts, lab members, collaborators, the UO Linguistics faculties and staffs, the undergraduate annotators who contributed to this project, and everyone who cheered me on behind the scenes, including my family and friends back in Korea. These individuals generously shared their time with me in ways both big and small. One of my favorite Korean film critics once said, "Love is presenting someone with your time." If that is a valid operational definition, then I am grateful to have been surrounded by such people.

Finally, I acknowledge the institutional support that made this work possible. I was fortunate to receive a GE-ship from the Department of Linguistics, which supported me throughout my graduate studies. The Raymund Fellowship, funded by UO alumnus Steve Raymund, enabled me to focus entirely on my research during my first year. In my final year, the Harold Gulliksen Psychometric Research Fellowship from ETS provided crucial support that allowed me to complete this dissertation. I am sincerely thankful for these opportunities.

TABLE OF CONTENTS

ACKNOWLEDGMENTS 4

TABLE OF CONTENTS 5

LIST OF FIGURES 9

LIST OF TABLE 10

1 Introduction 12

1.1 Background 12

1.2 Research focus and questions 16

1.3 Overview 17

2 Literature review 20

2.1 Linguistic complexity in L2 research: Definition and scope 20

2.2 Usage-based perspective on L2 development 21

2.3 Operationalizing complexity 23

2.3.1 Length-based syntactic complexity 23

2.3.2 Fine-grained syntactic complexity indices 23

2.3.3 Construction diversity: Focusing on ASCs 25

2.3.4 Lexical diversity 26

2.4 Operationalizing sophistication 27

2.4.1 Construction frequency and verb association strength 27

2.4.2 *n*-gram association strength 29

2.4.3 Lexical sophistication 29

2.5 Other applications of ASC/VAC indices 30

2.6 Extraction of ASCs: A methodological review 31

2.6.1 Manual approach 32

2.6.2 Automated approaches based on syntactic or semantic frames 33

2.6.3 Toward a PLM-enhanced approach: Bridging manual and automated approaches...36
2.7 Summary...38
3 Methodology...40
3.1 Tool design...40
3.2 Dataset construction...41
3.2.1 Annotation scheme...41
3.2.2 Annotated datasets...44
3.2.3 Annotation guidelines...45
3.2.4 Annotation process and review...46
3.3 ASC tagger development...48
3.3.1 Training setup...48
3.3.2 Model evaluation setup...49
3.3.3 Model evaluation result...50
3.4 ASC analyzer development...51
3.5 Summary...55
4 Measuring lexicogrammatical complexity and sophistication in L2 speaking assessment...57
4.1 Method...58
4.1.1 Datasets...58
4.1.2 Target constructs and indices...61
4.1.3 Statistical analyses...64
4.2 Results...66
4.2.1 RQ 1: Relationship between ASC-based indices and L2 oral proficiency scores...66
4.2.2 RQ 2: Extent to which ASC-based indices predict L2 oral proficiency scores...70
4.2.3 RQ 3: Extent to which ASC-based and other indices predict L2 oral proficiency scores...72

4.3 Discussion 75
4.3.1 RQ 1: Relationship between ASC-based indices and L2 oral proficiency scores 75
4.3.2 RQ 2: Extent to which ASC-based indices predict L2 oral proficiency scores 82
4.3.3 RQ 3: Extent to which ASC-based and other indices predict L2 oral proficiency scores 83
4.4 Summary 84
5 Measuring lexicogrammatical complexity and sophistication in L2 writing assessment 86
5.1 Method 87
5.1.1 Datasets 87
5.1.2 Additional construct and indices: Syntactic complexity 90
5.1.3 Statistical analyses 91
5.2 Results 92
5.2.1 RQ 1: Relationship between ASC-based indices and L2 writing proficiency scores 92
5.2.2 RQ 2: Extent to which ASC-based indices predict L2 writing proficiency scores 96
5.2.3 RQ 3: Extent to which ASC-based and other indices predict L2 writing proficiency scores 97
5.3 Discussion 102
5.3.1 RQ 1: Relationship between ASC-based indices and L2 writing proficiency scores 102
5.3.2 RQ 2: Extent to which ASC-based indices predict L2 writing proficiency scores 108
5.3.3 RQ 3: Extent to which ASC-based and other indices predict L2 writing proficiency scores 109
5.4 Summary 110
6 Conclusion 112
6.1 Summary of findings 112
6.1.1 Relationship between ASC-based indices and human-rated L2 proficiency 113
6.1.2 Extent to which ASC-based indices predict human-rated L2 proficiency 114

6.1.3 Extent to which ASC-based and other indices predict human-rated L2 proficiency114

6.2 Summary of tool and research applications 115

6.3 Implications.....115

6.4 Limitations 117

References.....119

Appendix.....143

A. Hyperparameter settings for the transformer-based NER model 143

B. JLE scoring rubric 144

C. ELLIPSE scoring rubric 145

LIST OF FIGURES

Figure 1. Overview of the development of ASC analyzer 41
Figure 2. Example of the annotation unit. 42
Figure 3. An example of ASC annotation in the CoNLL-U format 44
Figure 4. Evaluation of inter-annotator agreement 48
Figure 5. Categorizations of ASC-based indices 51
Figure 6. Distribution of ASCs in the JLE (learner) and SUBTLEX-US (reference) corpora 60
Figure 7. Correlations between L2 oral proficiency scores and selected indices of ASC use 67
Figure 8. Actual L2 oral proficiency scores vs. scores predicted by the best model (ASC-based indices). 72
Figure 9. Actual L2 oral proficiency scores vs. scores predicted by the best model (ASC-based and other lexicogrammatical indices) 74
Figure 10. Summary of the relative importance of each predictor in the best model (ASC-based and other lexicogrammatical indices) 75
Figure 11. Distribution of proportion indices across oral proficiency levels (x-axis) with individual y-axis scales for each index 78
Figure 12. Unified scale comparison of proportion indices across oral proficiency levels 78
Figure 13. Distribution of SOA indices (*DeltaPStructureCue*) across oral proficiency levels with individual y-axis scales for each index 81
Figure 14. Distribution of SOA indices (*DeltaPLemmaCue*) across oral proficiency levels (x-axis) with individual y-axis scales for each index 82
Figure 15. Distribution of ASCs in the ELLIPSE (learner) and EnCOW (reference) corpora 89
Figure 16. Correlations between L2 writing proficiency scores and selected indices of ASC use 92
Figure 17. Actual L2 writing proficiency scores vs. scores predicted by the best model (ASC-based indices) 97
Figure 18. Actual L2 writing proficiency scores vs. Scores predicted by the best model (ASC-based and other lexicogrammatical indices) 100
Figure 19. Summary of the relative importance of each predictor in the best model (ASC-based, lexicogrammatical, syntactic complexity indices) 102
Figure 20. Distribution of proportion indices across proficiency score groups (x-axis) with individual y-axis scales for each index 105
Figure 21. DeltaStructureCue of the top 50 verbs in the transitive simple construction 108

LIST OF TABLES

Table 1. Example of English ASCs 22
Table 2. Overview of previous studies on ASC/VAC production in L2 corpora 31
Table 3. Target ASCs and semantic-syntactic representations 43
Table 4. ASCs distribution in the gold-standard ASC treebank 47
Table 5. Distribution of ASCs in the gold-standard treebank across different domains 49
Table 6. F1 scores across ASC types, models, and domains 50
Table 7. Number of learners and tokens across L2 oral proficiency levels 59
Table 8. Descriptive statistics: Diversity indices 67
Table 9. Correlations between diversity indices and L2 oral proficiency scores 67
Table 10. Descriptive statistics: Proportion indices 68
Table 11. Correlations between proportion indices and L2 oral proficiency scores 68
Table 12. Descriptive statistics: Frequency indices 69
Table 13. Correlations between frequency indices and L2 oral proficiency scores 69
Table 14. Descriptive statistics: SOA indices 70
Table 15. Correlations between SOA indices and L2 oral proficiency score and SOA indices 70
Table 16. Linear model predicting L2 oral proficiency scores using the selected ASC-based indices 71
Table 17. Linear model predicting L2 oral proficiency scores using the selected ASC-based and other lexicogrammatical indices 72
Table 18. Frequency of ASC and ASC-verb combinations from the SUBTLEX-US corpus 79
Table 19. Summary the target constructs and related indices in measuring L2 writing proficiency 91
Table 20. Descriptive statistics: Diversity indices 93
Table 21. Correlations between diversity indices and L2 writing proficiency scores 93
Table 22. Descriptive statistics: Proportion indices 94
Table 23. Correlations between proportion indices and L2 writing proficiency scores 94
Table 24. Descriptive statistics: Frequency indices 95
Table 25. Correlations between frequency indices and L2 writing proficiency scores 95
Table 26. Descriptive statistics: SOA indices 95
Table 27. Correlations between SOA indices and L2 writing proficiency scores 96
Table 28. Linear model predicting L2 writing proficiency scores using the selected ASC-based indices 96

Table 29. Comparison of adjusted R^2 and incremental gains by included indices 98

Table 30. Linear model predicting L2 writing proficiency scores using ASC-based and syntactic complexity indices 98

Table 31. Linear model predicting L2 writing proficiency scores using ASC-based and other lexicogrammatical indices 99

Table 32. Linear model predicting L2 writing proficiency scores using ASC-based, lexicogrammatical, syntactic complexity indices 101

Table 33. Frequency of ASC and ASC-verb combinations from the EnCOW-US corpus 105

Table 34. SOA (*MI*) of ASC-verb combinations from the EnCOW-US corpus 107

1 Introduction

1.1 Background

Proficient second language (L2) users gradually develop the ability to produce lexicogrammatically complex and sophisticated language. While a simple sentence (e.g., "He left the room") conveys the same core meaning as a more refined alternative (e.g., "He promptly excused himself from the room"), the latter demonstrates greater linguistic complexity and sophistication. Generally, L2 users are expected to gain the skill to integrate such nuanced constructions into their communication as they develop (Gass et al., 2013).

Recognizing this progression, L2 researchers have commonly evaluated language production through the lens of complexity. Complexity gauges how extensively L2 users employ advanced linguistic forms, and these features serve as key indicators of higher-level language use, linking empirical observations to theoretical claims about linguistic development (Housen et al., 2012; Housen & Kuiken 2009; Norris & Ortega, 2009). Accordingly, a large number and variety of syntactic complexity indices have been proposed and evaluated (e.g., Biber et al., 2011; Lu, 2011; McNamara et al., 2014; Bulté & Housen, 2014).

However, researchers sometimes have faced both reliability and validity challenges when quantifying syntactic complexity (Kyle, 2016). From a reliability standpoint, the proposed indices often lacked consistency in their basic definitions (e.g., what constituent a "clause" varies across studies), as illustrated by varying approaches in Bardovi-Harlig and Bofman (1989), Polio (1997), and Wolfe-Quintero et al. (1998). These inconsistencies have extended to the level of granularity at which measures are applied. For instance, some studies counted each subordinate clause within a complex sentence as a separate unit, whereas others treated the entire embedded

structure as one clause (Larsen-Freeman, 2009; Norris & Ortega, 2009; Wolfe-Quintero et al., 1998), complicating comparisons across studies.

Equally important are validity concerns: while some indices (e.g., mean length of T-unit) reliably correlated with L2 writing proficiency in certain empirical work (Lu, 2011), they often failed to capture the emergence of specific syntactic constructions that signify advancing proficiency (Biber et al., 2011; Norris & Ortega, 2009). Moreover, many syntactic complexity measures focused exclusively on the length of syntactic units, which overlooked a growing body of usage-based evidence that language develops through the dynamic interplay of grammar and lexis (Ellis & Ferreira-Junior, 2009b; Römer, 2009).

To address these gaps, Kyle (2016) pioneered syntactic sophistication indices that gauge both the frequency of verb-argument constructions (VACs) and the strength of their association with corresponding verbs. Building on this operational definition, he developed computational tools to measure these indices and evaluated them in relation to L2 writing proficiency. In addition to Kyle's approach, subsequent studies have examined syntactic sophistication and constructional diversity in both spoken and written L2 production (Choi & Sung, 2020; Hwang & Kim, 2023; Kim & Ro, 2023; Kyle, 2016; Kyle & Crossley, 2017; Kyle et al., 2021). In these studies, methodologically, there have been two main approaches to extract target construction types from learner (and reference corpora). First, some studies have employed manual identification which relied on human judgment (often researchers themselves) to pinpoint target constructions. Second, recent research has employed dependency parsers and taggers to extract verb and arguments (e.g., using a syntactic pattern of *subject*-VERB-*indirect_object*-*direct_object* to extract *ditransitive* construction), leveraging syntactic frames for large-scale analyses of lexicogrammatical sophistication.

While the automatic extraction of target constructions represents significant progress in L2 corpus research, three major limitations remain. First, syntactic-frame approaches have not considered the semantic roles of arguments (e.g., agent, patient, goal, or result; Fillmore, 1968; Palmer et al., 2005), which are critical for capturing essential event meanings. Theoretically, constructions themselves are understood as conventional form-*meaning* mappings (Fillmore, 1988; Goldberg, 1995). They range from morphemes (e.g., past-tense suffix "-ed") and lexical items (e.g., "apple", "friend") and to more abstract patterns such as clausal-level templates (e.g., conditional patterns such as "if X, then Y") and argument structures (e.g., transitive constructions that imply an action performed on a theme). Consequently, overlooking semantic roles can obscure nuanced distinctions between constructions. For example, consider how an oblique argument shifts in interpretation: in "He put the vase on the shelf," the oblique phrase "on the shelf" functions as a core argument integral to the meaning of put, whereas in "He read the book in the library," the oblique "in the library" serves as an adjunct, providing contextual information. Incorporating semantic roles helps differentiate these constructions by clarifying which arguments are core and which are peripheral, particularly those expressing goals or results.

Second, although Kyle and Crossley (2017) demonstrated an important finding that less frequent and more strongly associated verb-VAC combinations reliably predict L2 writing proficiency, their method treated minor syntactic variations as entirely separate constructions. For example, the verb "have" can occur both in a simple transitive frame (*subject*–VERB–*direct_object*) and in a subordinate frame (*subordinator*–*subject*–VERB– *direct_object*), yet both realize the same core meaning. This procedure inflates the inventory of VAC types and obscures the central "have + object" pattern. From a usage-based constructionist perspective, such variants

are simply surface forms of the same basic argument structure construction (ASC)—for instance, the transitive construction (Hwang & Kim, 2023).[1] A more coherent methodology would first identify the major clause patterns (e.g., the transitive ASC) and then subdivide them according to the presence versus absence of a subordinator. This hierarchical strategy, advocated by corpus-based grammarians (Biber et al., 1999; Quirk et al., 1985) and recently applied by Park and Sung (2024), provides a more structured, theoretically grounded framework than an undifferentiated bottom-up enumeration of every minor variant.

Third, recent advances in large pre-trained language models (PLMs; e.g., BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019]) underscore transformative potential of fine-tuning language models dedicated to linguistic structure analysis. While PLMs have already advanced traditional linguistic tasks such as dependency parsing (Honnibal et al., 2020; Van Nguyen et al., 2021) or semantic role labeling (Shi & Lin, 2019), emerging research now leverages their latent linguistic "knowledge" to address more nuanced linguistic phenomena, including rhetorical stance features (Eguchi & Kyle, 2023), abstract meaning representations (Ettinger et al., 2023), and pragmatic discourse markers such as apologies (Yu et al., 2023). By leveraging contextualized embeddings and self-attention mechanisms, PLMs may provide fine-grained disambiguation of syntactic and semantic relationships within L2 corpora. Such capabilities would be able to support more precise and scalable classification of ASCs by systematically disentangling core arguments (e.g., agent, patient) from peripheral adjuncts (e.g., locative phrases), thereby addressing the limitations of traditional syntactic-frame analyses.

[1] By "VACs," I refer to the pairing of "a verb slot and the arguments it takes" (Kyle, 2016). This concept aligns closely with ASCs; however, VACs are typically framed around specific verbs and may include an unrestricted set of clausal-level arguments, reflecting a bottom-up approach. In contrast, ASCs focus on a more constrained set of syntactic and semantic relationships among arguments in a clause, adopting a top-down perspective. The specific set of ASCs used in this study are discussed in detail in the following sections.

In summary, L2 users are expected to gradually acquire advanced lexicogrammatical constructions, and researchers often track this development through measures of syntactic complexity. However, traditional complexity metrics have suffered from reliability issues and often overlooked the subtle lexicogrammatical patterns that mark high L2 proficiency. To address this, Kyle (2016) introduced syntactic sophistication indices that quantify VAC frequency and the strength of verb-argument associations. Nevertheless, three gaps remain. First, some measures fail to encode semantic roles, making it hard to distinguish constructions that differ only in core versus peripheral arguments. Second, they over-split minor syntactic variants, obscuring the central patterns of productive proficiency. A more systematic framework for identifying core ASCs would enhance both theoretical clarity and the comparability of empirical studies. Finally, leveraging PLMs to distinguish core arguments from peripheral adjuncts could further refine this approach.

1.2 Research focus and questions

This study develops and evaluates ASC-based indices for assessing lexicogrammatical complexity and sophistication in L2 speaking and writing assessments. First, an ASC tagger is developed by fine-tuning RoBERTa on a gold-annotated treebank of nine target ASC types to create an ASC tagger. Next, an ASC analyzer, a computational tool that calculates ASC usage across 50 indices, is built. These indices include three text-internal diversity measures (e.g., the moving-ratio type-token ratio of ASCs) and nine proportion measures (e.g., the proportion of passive ASCs per text). They also incorporate two text-external frequency measures (e.g., normalized average frequency of verb-ASC combinations against reference corpora) and thirty-six strength-of-association measures (e.g., mutual information scores between verbs and caused-

motion ASCs). In addition, in line with research emphasizing L2 productive proficiency as a multifaceted construct (e.g., Bulté & Roothooft, 2020; Eguchi & Kyle, 2020; Kim et al., 2018; Kyle & Eguchi, 2021, 2023; Saito, 2020), ASC-based indices are examined with other lexicogrammatical and syntactic measures to assess the unique variance they explain. The study addresses the following research questions (RQs):

RQ 1. What is the relationship between ASC-based indices and human-rated L2 proficiency scores in L2 speaking and writing assessments?

RQ 2. To what extent do ASC-based indices independently predict L2 proficiency scores?

RQ 3. How much additional predictive power do ASC-based indices provide when combined with previously established lexicogrammatical and syntactic measures?

1.3 Overview

The dissertation is organized into six chapters that collectively explore how ASC-based indices can operationalize and measure lexicogrammatical complexity and sophistication in L2 speaking and written assessments, and how these indices enable fine-grained analysis of L2 productions across different proficiency scales. Chapter 1 introduced the study's aims, situated ASC-based measures within the broader landscape of L2 complexity research, and presented the research questions.

Chapter 2 reviews the literature on linguistic complexity and sophistication, focusing on the various measures used to operationalize these constructs in L2 research. It then addresses key methodological considerations, including several approaches that have been explored for extracting ASCs—ranging from manual methods to those leveraging syntactic and semantic

information—and highlights promising results from recent studies using PLMs to extract target linguistic features with high accuracy.

Chapter 3 details the methodology. It first describes the tool's overall design and the construction of the annotated dataset, covering the annotation scheme, guidelines, and review process. It then explains the development of the ASC tagger, outlining model-training and evaluation protocols. It concludes with the ASC analyzer, which calculates a range of ASC-based indices capturing diversity, proportion, frequency, and strength of association between ASCs and verb lemmas. This work was published in *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)* in March 2024 as *Annotation Scheme for English Argument Structure Constructions Treebank*, and in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* in November 2024 as *Leveraging Pre-trained Language Models for Linguistic Analysis: A Case of Argument Structure Constructions*. Hakyung Sung and Dr. Kris Kyle were the principal investigators for this work.

Chapter 4 reports a corpus-based investigation of L2 speaking assessment, examining the relationship between ASC-based indices and human-rated oral proficiency scores. These indices are evaluated in conjunction with previously established measures of lexicogrammatical complexity and sophistication using a multivariate regression framework to assess their additional predictive value. This work has been accepted for publication in *Studies in Second Language Acquisition* (in press) as *Usage-based Analysis of L2 Oral Proficiency: Characteristics of Argument Structure Construction Use*. Hakyung Sung and Dr. Kris Kyle were the principal investigators for this work.

Chapter 5 shifts the focus to L2 writing assessment, offering parallel analyses of the ASC-based indices. It evaluates their contribution to explaining variance in writing proficiency scores when combined with conventional lexicogrammatical and syntactic measures.

Chapter 6 summarizes the main findings and contributions of the study, highlighting both methodological and theoretical implications. It also addresses limitations, offer directions for future research, and discusses the broader potential of ASC-based analyses for more sophisticated understanding of L2-English (or, more broadly, English) production.

2 Literature review

2.1 Linguistic complexity in L2 research: Definition and scope

In L2 research, *complexity* has been widely acknowledged as an important construct, yet consensus on its definition remains elusive[2] (Hinkel, 2003). Broadly speaking, the construct comprises two intertwined dimensions: *absolute complexity* and *relative complexity* (Bulté & Housen, 2012; Bulté et al., 2024). Absolute complexity focuses on formal characteristics of language. In the case of syntactic structures, complexity increases with the number of discrete units and the depth of their hierarchical embedding. This principle extends to other linguistic levels as well. Lexical complexity, for instance, can be assessed by the range of vocabulary in a text or by the structural richness of individual words, such as the number of morphological derivations associated with a given lexical item.

Relative complexity reflects the cognitive effort required to learn, process, and deploy particular forms. Drawing on studies of lexical development (Laufer & Nation, 1995; Linnarud, 1986), Kyle (2016) extended the notion of *sophistication* to absolute complexity, coining the term *syntactic sophistication*. Infrequent or highly contingent constructions are considered more sophisticated because they are more likely to be produced by advanced L2 users than by beginner L2 learners. Longitudinal evidence supports a recurring trajectory in which L2 learners initially rely on prefabricated but structurally dense chunks (e.g., those involving embedded syntactic structures), then gradually decompose them into more flexible and generative patterns.

[2] According to Hinkel (2003), numerous and often competing definitions may exist for linguistic simplicity and complexity. Some definitions on simplicity rely on frequencies of syntactic structures and words to identify the most common and useful for learners (Nation, 1990), while others emphasize the minimal grammar and lexis needed to express ideas (Allen, 1983). Still others focus on the simplicity or complexity of syntactic and morphological derivations (Biber et al., 1999; Quirk et al., 1985). Each of these approaches has its strengths and weaknesses, particularly because they are often based on language produced by native English speakers rather than on L2 in specific contexts. In this study, the conceptualization proposed by Bulté and Housen (2012) is highlighted to illustrate a well-known framework for operationalizing and measuring linguistic complexities in L2 production research.

As their proficiency increases, they eventually reintroduce low-frequency, semantically specialized constructions—this time used productively—as part of their expanding expressive repertoire (Myles et al., 1999). Cross-linguistic proximity can influence this timeline: learners whose L1 is typologically closer to English (e.g., Dutch or German) attain near-native control of complex phrasal verbs earlier than learners from more distant L1s (e.g., Spanish or Korean) (Hawkins & Buttery, 2010; Kim et al., 2020; Sung, 2019).

The boundary between absolute and relative complexity may not be always clear-cut, because a structure may consist of numerous parts (which indicates high absolute complexity), while simultaneously posing processing difficulty for L2 learners (which involves relative complexity). For terminological clarity, the present study uses *complexity* to refer to research concerned with the number and diversity of linguistic structures, and *sophistication* to refer to relative complexity as operationalized from a usage-based perspective involving relative frequency and contingency.

2.2 Usage-based perspective on L2 development

From a usage-based constructionist perspective, language is understood as a dynamic network of constructions—form-meaning pairings that become entrenched through repeated exposure (Fillmore et al., 1988; Goldberg, 1995; Langacker, 1987). Constructions emerge from actual language use, both receptive (listening, reading) and productive (speaking, writing), and are shaped by the frequency, distribution, and co-occurrence patterns of the input (Bybee, 2010; Diessel, 2015; Ellis, 2012). Because learning is usage-driven, linguistic knowledge is assembled incrementally, not imposed by static, decontextualized rules. Each learner therefore develops a lexicogrammatical repertoire that reflects their unique history of language experience (Ellis &

Ferreira-Junior, 2009a, b; Gries & Wulff, 2005; Herbst, 2016; Hwang & Kim, 2023; Kim & Ro, 2023; Kyle, 2016; Römer et al., 2014).

Empirical studies grounded in this perspective investigate constructions at multiple levels of granularity. At the clausal level, ASCs have received particular attention for their foundational role in clause formation. ASCs are defined as abstract, schematic pairings of form and meaning that encode core semantic patterns such as motion, causation, and transfer (Goldberg, 1995). They represent "a special subclass of constructions that provides the basic means of clausal expression in a language" (Goldberg, 1995, p. 3). These constructions emerge when speakers repeatedly combine simpler symbolic units (e.g., lexical items) into stable higher-level grammatical patterns, a process of gradual entrenchment that transforms into generalized grammatical knowledge (Langacker, 1990). As language users accumulate experience, their ASC repertoire expands in both complexity and sophistication. Rooted in frame semantics (Fillmore, 1976), ASCs are typically distinguished by the number and nature of their obligatory semantic arguments (Diessel, 2004). Examples of English ASCs are provided in Table 1.

Table 1. Example of English ASCs

ASC	**Meaning**	**Example**
Intransitive-motion	X moves $Y_{path/location}$	*The fly buzzed into the room.*
Intransitive-result	X become Y_{state}	*The river froze solid*
Transitive	X acts on Y	*Tom kicked the ball.*
Caused-motion	X causes Y to move Z_{goal}	*Tom kicked the ball into the net.*
Ditransitive	X causes Y to receive Z	*Tom kicked me the ball.*
Transitive-result	X causes Y to become Z_{state}	*Tom kicked the ball flat.*

While ASCs are central to clause-level expression, usage-based theories emphasize that constructions span multiple linguistic layers—from single words and multiword sequences to complex syntactic templates (Goldberg & Suttle, 2010). Given this, a comprehensive analysis of L2 development should move beyond isolated grammatical or lexical items and instead adopt a

multivariate lens. This perspective aligns with recent research that models L2 productive proficiency as a multivariate construct (e.g., Eguchi & Kyle, 2020; Kyle & Eguchi, 2021, 2023; Nation, 2001; Römer, 2009).

2.3 Operationalizing complexity

Previous studies have operationalized syntactic and lexicogrammatical complexity using a range of measurement unit designed to capture both the length and variety of linguistic structures. This section organizes these approaches into four categories: length-based syntactic complexity, fine-grained syntactic complexity, construction diversity, and lexical diversity.

2.3.1 Length-based syntactic complexity

Traditional structural approaches quantify syntactic complexity by measuring the average length of production units such as clauses, sentences, or T-units. Among these, length-based indices such as mean length of clause (MLC) and mean length of T-units (MLTU) remain highly influential. According to Bulté and Housen (2012, pp. 36-37), their popularity stems from three key reasons: they encapsulate syntactic diversity and depth, integrate both phrasal and clausal structures, and align with long-standing traditions in L1 and L2 acquisition research (e.g., Hunt, 1965; Ortega, 2003; Wolfe-Quintero et al., 1998). Subsequent empirical studies of L2 writing have consistently demonstrated that these measures (e.g., MLTU) correlate positively with L2 proficiency (e.g., Lu, 2011; Hwang et al., 2020; Kim, 2014).

2.3.2 Fine-grained syntactic complexity indices

Despite the widespread use of length-based syntactic complexity indices, the theoretical validity of those measures remains contested. Scholars have long called attention to the lack of a robust theoretical framework for using these indices to fully capture L2 syntactic development

(Bardovi-Harlig, 1992; Biber et al., 2011, 2016; Bulté & Housen, 2012; Norris & Ortega, 2009). For example, Biber et al. (2011) argued that "there is little empirical evidence that T-unit measures and dependent clause measures are appropriate for the assessment of writing development." To illustrate, they compared a conversational and an academic text that share the same T-unit lengths but differ substantially in how they achieve complexity. The conversational text relies primarily on subordination, whereas the academic text employs elaborate noun-phrase constructions, suggesting that T-unit metrics may underrepresent advanced literacy skills. Similarly, Larsen-Freeman (2009) and Norris and Ortega (2009) cautioned that an exclusive focus on clausal or T-unit indices can obscure important linguistic variation, resulting in oversimplified interpretations of L2 production. Accordingly, Norris and Ortega (2009) advocated for "developmentally sensitive and interlanguage-oriented measures" (p. 574) to provide a more nuanced account of L2 syntactic development. In response to these concerns, researchers have begun to explore additional or alternative indices of syntactic complexity that move beyond coarse-grained, length-based constructs to target more fine-grained syntactic features. For instance, Kyle (2016) introduced fine-grained measures of clausal complexity (e.g., nominal subjects per clause) and phrasal complexity (e.g., dependents per nominal), demonstrating how these metrics correlate with L2 writing proficiency. This approach resonates with Biber et al.'s (2011) observations on academic writing, which showed that advanced L2 writers tended to produce more complex noun phrases and frequent nominal modifications. By capturing these subtler structural dimensions, fine-grained syntactic complexity indices have offered a more developmentally sensitive and comprehensive view of structural features, particularly those that traditional length-based measures might overlook in advanced L2 writing.

2.3.3 Construction diversity: Focusing on ASCs

Shifting the focus from the formal characteristics of syntax to specific grammatical constructions, a growing body of research has highlighted the developmental trajectory of ASCs (or more broadly VACs), as a marker of L2 proficiency from a usage-based constructionist perspective (e.g., Choi & Sung, 2020; Hwang & Kim, 2023; Kim & Ro, 2023; Liu & Lu, 2024; Park & Sung, 2024). These studies hypothesized that L2 learners expand their ASC inventories over time, moving from simpler forms (e.g., simple transitive) to more complex constructions involving transfer (e.g., ditransitive), motion (e.g., caused-motion), and voice distinctions (e.g., passive). This emphasis on ASC complexity aligns with Goldberg's (1999) construction grammar framework, which organizes constructions in an inheritance hierarchy by their number of semantic roles: for example, at the top sit the simple subject-predicate constructions, followed by two-place constructions that introduce a theme (*agent*-VERB-*theme*), and further down those that add a goal (*agent*-VERB-*theme*-*goal*).

In relation to L2 writing production, Hwang and Kim (2023) showed that more proficient L2 writers use a higher proportion of these complex ASCs. Similarly, Kim et al. (2023) compared ASC-based measures to traditional T-unit indices in predicting Korean EFL learners' writing proficiency, finding that complex constructions (i.e., resultative, caused-motion, and periphrastic causative) serve as stronger proficiency markers.

Although evidence from the L2 speaking domain is more limited, it points in a similar direction. Choi and Sung (2020), focusing on fluency in English communication (Fillmore, 1979; Faerch et al., 1984), showed that ASCs (particularly simple transitive constructions) accounted for most of the variance in learner fluency. They found that while word-level measures had minimal impact, sentence-level utterances defined by ASC types significantly distinguished

fluency among Korean middle-school learners. Kim and Ro (2023) also found that college-level speakers, particularly advanced learners, produced a more diverse range of verb-construction combinations, indicating greater flexibility and lexical specificity in their use of ASCs.

Meanwhile, as ASCs surface through specific verb fillers, several studies have treated VACs as a "structured inventory" that evolves over time. Park and Sung (2024), for example, mapped the order in which VAC types typically emerge and showed how they aggregate into functional repertoires. Liu and Lu (2024) likewise reported that advanced writers command a wider VAC range and that their distribution more closely mirrors native benchmarks.

2.3.4 Lexical diversity

At the word level, complexity has been often operationalized through lexical diversity, which refers to the range of distinct word types a learner uses within a given text.[3] Lexical diversity has traditionally been measured by calculating the type-token ratio (TTR), which expresses the proportion of different word forms (types) relative to the total number of words (tokens) in a text. However, because TTR is highly sensitive to text length, more sophisticated measures have been developed to minimize this dependency. One of the most widely used and robust is the moving-average type-token ratio (MATTR) (Bulté et al., 2024, p. 17).

MATTR estimates lexical diversity by calculating the type-token ratio within a fixed-size moving window across a text and then averaging these values to produce a stable score. Unlike traditional measures such as TTR (Johnson, 1944) and rootTTR (Guiraud, 1960), which are highly sensitive to text length, MATTR offers more reliable comparisons across texts of different

[3] According to Jarvis (2013), lexical diversity itself is a *multidimensional* construct, which includes volume (the total number of word tokens), abundance (the number of unique word types), variety (the proportion of unique words), evenness (the distribution uniformity of unique words), disparity (the semantic relatedness of words), specialness (the presence of specific words perceived as enhancing diversity), and dispersion (the size of intervals between repetitions of the same words). In this study, attention is restricted to the *variety* dimension, and I set out the specific indices used to measure it.

lengths. When the window size is optimized for specific contexts (e.g., essay writing, spontaneous speech; cf. Kyle et al., 2024), MATTR has demonstrated strong validity in capturing lexical diversity in both written (Zenker & Kyle, 2021) and spoken (Kyle et al., 2024) L2 production.

2.4 Operationalizing sophistication

Sophistication reflects the relative difficulty and advanced use of target constructions. Constructions that are infrequent in the input or strongly associated with specific lexical items have been considered more sophisticated, as they are more likely to be produced by advanced L2 users (Kyle, 2016). This section summarizes previous studies across three categories: construction frequency and verb association strength, *n*-gram association strength, and lexical sophistication.

2.4.1 Construction frequency and verb association strength

Usage-based approach views grammar and lexis as a continuum of form-meaning mappings (Bybee, 2010; Ellis & Wulff, 2014). Treating them in isolation obscures an important source of functional complexity: a single verb can participate in multiple ASCs and convey different meanings, while a single construction can host a wide range of lexical fillers. Analyses that integrate the two therefore provide a more sophisticated picture of how advanced L2 learners express meaning in context.

Previous studies have pointed out that early ASC development is grounded in high-frequency, prototypical verbs that strongly cue particular constructions (Clark, 1987; Goldberg et al., 2004; Ninio, 1999; Viberg, 2002). Ellis and Ferreira-Junior (2009b) showed that verbs such as “go” (for intransitive motion construction), “put” (caused-motion), and “give” (ditransitive)

serve as entry points: they enable L2 learners to produce the target constructions and, over time, to diversify their semantic verb choices. Similarly, Kim and Rah (2016) found that light verbs (e.g., "do," "make," "have") serve as important facilitators for lower-proficiency learners, while more advanced writers move beyond light verbs to a broader, semantically richer verb inventory.

Subsequent studies have moved beyond learner-input comparisons to benchmark learner production against advanced-speaker norms drawn from large reference corpora. By leveraging relative frequency (i.e., the prevalence of an ASC or ASC-verb combination) and strength of association (SOA) (i.e., the statistical contingency between a verb and its construction), these studies gauge L2 users' lexicogrammatical sophistication (e.g., Kim & Ro, 2024; Kyle, 2016; Kyle & Crossley, 2017; Kyle et al., 2021; Liu & Lu, 2024; Mostafa & Crossley, 2020; Römer & Berger, 2019). This focus on frequency and contingency has been also theoretically motivated. For example, it aligns with Goldberg's (1999, ch. 2) distinction between *argument roles* (i.e., the slots specified by a construction) and *participant roles* (i.e., the semantic functions supplied by the verb. Tracking how often L2 users pair particular verbs with particular constructions thus reveals how they progressively map participant roles onto argument structures. In addition, Gries and Ellis (2015) suggested that stronger form-function mappings increase the reliability with which learners choose constructions, thereby facilitating both comprehension and production.

In empirical studies, Kyle and Crossley (2017) found that human-rated essay quality was negatively correlated with the relative frequency of verb-VAC combinations but positively correlated with their SOA, suggesting that higher L2 writing proficiency is linked to the use of less common yet more tightly bounded verb-VAC combinations. Likewise, Römer and Berger (2019) found that as German and Spanish learners' proficiency increased, their use of VACs

(which was measured through measures of frequency and SOA) approached that of advanced speakers' norms.

2.4.2 *n*-gram association strength

At the multiword level, lexicogrammatical sophistication has also been operationalized through the relative frequency and association strength of *n*-gram collocations (e.g., Bestgen & Granger, 2014; Eguchi & Kyle, 2020; Garner et al., 2019; Kim et al., 2018; Kyle & Eguchi, 2023). Research shows that more proficient L2 users favor conventional bigrams and trigrams that occur frequently and exhibit strong lexical contingencies. Kim et al. (2018), for example, found that writers with higher lexical-proficiency scores produced a larger proportion of high-frequency bigrams and trigrams. Eguchi and Kyle (2020) also observed that higher oral proficiency ratings were tied to bigrams and trigrams with stronger statistical associations, as captured by measures such as mutual information (MI), T-scores, and delta P values.

Recent research has also expanded into *dependency* bigrams, examining bigrams in particular syntactic relations such as *noun-adjective*, VERB-*subject*, and VERB-*direct_object* combinations (Kyle & Eguchi, 2023). They found small to moderate correlations between dependency bigram indices and L2 oral proficiency scores, with more proficient L2 speakers using more strongly associated *noun-adjective* and VERB-*subject* combinations, findings consistent with recent research using L2 writing samples (e.g., Paquot, 2018; Rubin et al., 2021).

2.4.3 Lexical sophistication

At the word level, lexical sophistication is generally defined as the extent to which learners use relatively rare or advanced vocabulary in their texts (Laufer & Nation, 1995, Meara & Bell, 2001; Read, 2000). Three major constructs have been widely used. First, *frequency*-based sophistication gauges how often a word appears in a reference corpus; more advanced

writers draw on a broader range of low- to mid-frequency items that are appropriate to the domain (Crossley et al., 2009; Eguchi & Kyle, 2020; Kyle & Crossley, 2015; Laufer & Nation, 1995; Meara & Bell, 2001). Second, *concreteness* considers the abstractness of lexical choices, with higher proficiency linked to greater use of abstract vocabulary (Brysbaert et al., 2014; Crossley et al., 2011; Ellis & Beaton, 1993; Saito et al., 2016). Third, *contextual distinctiveness* refers to how broadly or narrowly a word is distributed across different phonological, lexical, or semantic contexts. Words that appear in a wide range of contexts tend to reflect general usage, whereas those that occur in fewer, more specific contexts are often considered more difficult or advanced (Berger et al., 2017). In L2 research, contextual distinctiveness has been operationalized as the distribution of a word across documents, and their use tends to increase with proficiency (Eguchi & Kyle, 2020; Kiss et al., 1973; Kyle & Crossley, 2015; Kyle et al., 2018; McDonald & Shillcock, 2001; Nelson et al., 2004).

2.5 Other applications of ASC/VAC indices

More recently, researchers have increasingly applied ASC/VAC measures as indices of lexicogrammatical complexity and sophistication to investigate how task design, genre, proficiency, and learning environment shape L2 writing. For example, Casal et al. (2022) used a genre-based approach to examine 11 VAC types in 400 research articles across engineering and social sciences, finding that certain VACs occur more often across social science fields while others remain discipline-specific. Similarly, Abdi Tabari et al. (2024) investigated how task complexity and repetition affect VAC use among 96 high-intermediate English-as-a-second language (ESL) students, showing that VAC frequency and sophistication capture nuanced variations across different writing tasks. Kim and Ro (2024) analyzed 3,196 L2 English essays

(narrative and argumentative) and found that genre and proficiency interact significantly to shape writers' lexicogrammatical choices (e.g., construction frequency/diversity, verb-construction combinations, verb-construction strength of associations), underscoring the role of context and learner experience in L2 writing development. Lastly, Li and Yu (2024) focused on argumentative essays by college-level Asian learners and found that higher-proficiency students, as well as those in ESL settings, produced more diverse and lower-frequency verb-construction combinations. These findings highlight the combined influence of proficiency and learning environment (ESL vs. English-as-a-foreign-language [EFL]) on VAC usage. Table 2 summarizes key studies discussed thus far, highlighting their domains and task types.

Table 2. Overview of previous studies on ASC/VAC production in L2 corpora

Study	**Domain**	**Task type**
Kyle & Crossley (2017)	Written	Task from TOEFL Public dataset (argumentative)
Römer & Berger (2019)	Written	Tasks from EFCAMDAT (mixed)
Mostafa & Crossley (2020)	Written	Descriptive writing
Kyle et al. (2021)	Written	General essays
Casel et al. (2022)	Written	Research articles
Hwang & Kim (2023)	Written	Task from YELC (argumentative)
Kim et al. (2023)	Written	Task from YELC (argumentative)
Kim & Ro (2024)	Written	Tasks from YELC (argumentative, narrative)
Li & Yu (2024)	Written	Tasks from ICNALE (argumentative)
Liu & Lu (2024)	Written	Tasks from EFCAMDAT2 (mixed)
Park & Sung (2024)	Written	Argumentative and narrative writing
Abdi Tabari et al. (2024)	Written	Argumentative writing
Choi & Sung (2020)	Spoken	Group conversations
Kim & Ro (2023)	Spoken	Task from YELC (argumentative)

Notes. EFCAMDAT (EF-Cambridge Open Language Database, Alexopoulou et al., 2015; Geertzen, 2013); YELC (Yonsei English Learner Corpus, Rhee & Jung, 2014); ICNALE (the International Corpus Network of Asian Learners of English, Ishikawa et al., 2011)

2.6 Extraction of ASCs: A methodological review

A major challenge in learner corpus research is the reliable identification of ASCs and linking each construction to its governing verbs (Kyle & Sung, 2023). Manual annotation has

long been the default approach, but recent studies have turned to computational approaches that leverage dependency parsing or semantic role labeling. This section reviews these approaches and discusses related work that leverages PLMs for high-accuracy feature extraction.

2.6.1 Manual approach

The manual approach to ASC extraction begins by defining target construction patterns according to a theoretical framework, then identifying matching instances in the corpus through human annotation (e.g., Choi & Sung, 2020; Kim & Rah, 2016; Park & Sung, 2024; Sung & Kim, 2022). For example, Kim and Rah (2016) manually coded four construction types (i.e., transitive, ditransitive, caused-motion, and resultative) in an L2 written corpus based on Goldberg's (1995) framework. Other studies have narrowed their focus to a predetermined list of candidate verbs, manually verifying each occurrence against the intended construction (e.g., Levin, 1993; Ellis & Ferreira-Junior, 2009a, b; Romain, 2022). Ellis and Ferreira-Junior (2009b), for instance, annotated L1 and L2 interview data (Perdue, 1993) for three construction types (i.e., VERB-*locative*, VERB-*direct_object-locative*, and VERB-*indirect_object-direct_object*) by first selecting verbs of interest and then conducting manual analyses.

A key strength of this approach lies in its internal reliability and transparency. Because annotations are applied according to predefined criteria, researchers can clearly justify their decisions and maintain consistency in how annotation schemes are used within a given study.

However, the approach presents notable limitations in terms of external reliability, replicability, and scalability. In terms of external reliability, it is often difficult for researchers outside the original research group to adopt the same annotation scheme, which complicates the generalizability of findings. This issue is amplified when analyses are based on small, manually

annotated learner corpora, making it challenging to extend conclusions to larger or more diverse datasets. In particular, such constraints limit meaningful comparisons with more proficient or native speaker data typically found in large reference corpora.

Replicability is also a concern. Despite the use of annotation guidelines, manual annotation is inherently subjective and influenced by researchers' interpretations. As a result, variations in how different annotators apply the same criteria can lead to inconsistencies across studies, reducing confidence in the reproducibility of results.

Finally, the scalability of this approach is limited due to its labor-intensive nature. A comprehensive examination of ASC usage requires identifying every relevant verb and construction in the corpus, a task that quickly becomes impractical as corpus size grows. Together, these challenges highlight the need for more efficient, generalizable methods of ASC extraction.

2.6.2 Automated approaches based on syntactic or semantic frames

2.6.2.1 Automatic extraction of constructions via syntactic frames

The use of syntactic dependency representations in treebanks and parsers (de Marneffe & Manning, 2008) has provided a basis for exploring automated approaches to ASC extraction. In early studies, O'Donnell and Ellis (2010), for example, used a dependency-parsed version of the BNC (Andersen et al., 2008) to test whether dependency tags could be used to extract target constructions. This method enabled the accurate extraction of some constructions (e.g., a VERB-*preposition-noun* construction such as "talked about it," Römer et al., 2014), yet the underlying parser's overall performance (F1 ≈ 0.76) was too modest for large-scale deployment.

Subsequent neural-network-based parsers (e.g., Chen & Manning, 2014) pushed dependency parsing accuracy much higher, supporting syntactic-frame based ASC annotations. Researchers extracted clause-level frames from these parses and mapped them to ASC categories (Hwang & Kim, 2023; Kyle, 2016; Kyle & Crossley, 2017). Syntax alone, however, often lacks the semantic resolution needed for reliable classification: one construction can surface in several frames, and one frame can encode multiple constructions. For example, the syntactic pattern *subject*-VERB-*object*-*oblique*$_{prep\_on}$ can represent both a simple transitive construction (e.g., "Subject found this object [on a bulletin board]") and a caused-motion construction (e.g., "Subject put it [on my hand]") (Kyle & Sung, 2023).

To add more cues, Hwang and Kim (2023) layered rule-based heuristics over dependency parses produce by *spaCy* (Honnibal et al., 2020). Their top-down decision tree used lexical cues (e.g., the expletive *there* to flag a *there*-construction) or syntactic cues (e.g., co-occurring direct and indirect objects to mark a ditransitive construction) to classify eleven ASC types. Although the system achieved an overall F1 of 0.82 (precision = 0.86; recall = 0.82), three issues constrain its broader use: First, it was unclear whether the system is designed to tag only finite clause types. Second, the system's ability to generalize to unseen syntactic structures or lexical items remains uncertain, particularly since the rules were manually derived. Third, although the authors report an overall F1 score, the evaluation lacks detailed analysis by construction type. It is therefore difficult to determine which constructions were classified with high accuracy and which were more error prone.

2.6.2.2 Automatic extraction of constructions via semantic frames

While earlier approaches primarily relied on syntactic frames to identify ASCs, recent studies have turned to semantic frames as an alternative means of capturing the underlying ASCs

(Kyle & Sung, 2023). Since ASCs consist of "argument roles" (Goldberg, 1995) that correspond to traditional semantic roles (e.g., agent, patient, theme, goal), using corpora annotated with semantic role labels (e.g., PropBank, Palmer et al., 2005; Universal Proposition treebank, Akbik et al., 2015) or employing automated semantic role labeling (SRL) systems (Gardner et al., 2018; Shi & Lin, 2019) can be useful for large-scale ASC extraction. This approach parallels earlier efforts that focused on syntactic cues, but it directly targets the semantic structure of clauses.

There are, however, two major challenges with relying solely on SRL approach for ASC extraction. First, state-of-the-art SRL systems still lack the accuracy required for reliable construction tagging, especially with learner language or other non-canonical input. To the best of current understanding, state-of-the-art SRL models typically achieve F1 scores in the mid-80s (e.g., 85 on CoNLL-2005; Shi & Lin, 2019). More recent evaluations similarly show that SRL systems continue to underperform compared to syntactic parsers, especially when applied to complex or diverse input (cf. recent semantic role labeling survey: Chen et al., 2025). In contrast, modern dependency parsers often reach labeled attachment scores above 90 (e.g., Dozat & Manning, 2016; Kulmizev et al., 2020), making them more reliable for structural analysis at scale. Second, the abstract argument labels used in many SRL systems (e.g., ARG0, ARG1 in the PropBank) do not consistently correspond to the theoretical roles that define ASCs. For example, in the sentence "She gave him a book," ARG0 might be assigned to "she", ARG1 to a "book," and ARG2 to "him", but this labeling does not necessarily lead to the conclusion that the construction is ditransitive (as ARG2 often could be completely different argument roles such as instrument or attribute; cf. Gildea & Jurafsky, 2002), which makes it difficult to map these roles to specific ASC types in a consistent manner.

One potential solution is to extract gold-annotated semantic roles from annotated corpora, group them into semantic frames, and then use linguistic knowledge to map each frame to the corresponding ASC types. These semi-automatically mapped ASCs can subsequently serve as training data for a sequential learning model, offering a more scalable yet linguistically grounded approach to construction identification. For example, Kyle and Sung (2023) combined resources such as the UP treebank (Akbik et al., 2015), VerbNet (Schuler, 2005), and FrameNet (Fillmore et al., 2003) to extract semantic roles and associated sentences from a subset of the English Web Treebank (Silveira et al., 2014). They grouped these semantic roles into frames and manually assigned ASC labels to each frame, creating a silver-annotated ASC treebank. A transformer model trained on these silver annotations, using RoBERTa embeddings, achieved an F1 of 0.918, outperforming models based solely on verb lemmas, dependency-based syntactic frames, or a combination of both. Even so, accuracy could rise further by incorporating additional gold-standard annotations, a direction the authors themselves highlight for future work (Kyle & Sung, 2023).

2.6.3 Toward a PLM-enhanced approach: Bridging manual and automated approaches

To briefly summarize, researchers have explored different methodologies for ASC identification. First, manual annotation has long been the standard for identifying ASCs and their associated verbs. Second, scalable computational approaches have leveraged syntactic or semantic frames, using dependency parses or semantic role labels drawn from linguistic databases or NLP tools. While these frame-based approaches markedly reduce labor, they still struggle with ambiguous constructions that fall outside the patterns captured by automatic extraction.

A promising next step is to train PLMs using gold-standard annotations, constructed through a dedicated ASC treebank, a strategy that revives the rigor of manual tagging while scaling it through automation. This approach enables the creation of high-quality datasets that can serve multiple purposes: (1) as training data for supervised learning tasks using encoder-only models, particularly for sequence-based named entity recognition (NER); (2) as in-context examples for few-shot learning with decoder-only models in prompting-based settings; (3) as test sets for robust model evaluation across different architectures.

Indeed, recent advancements have highlighted the potential of PLMs in automated linguistic annotation. Encoder-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), grounded in the Transformer architecture (Vaswani et al., 2017), have been extensively pre-trained on large text corpora and shown to capture morpho-syntactic and contextual sentence-level information in their embeddings (Hewitt & Manning, 2019; Miaschi & Dell'Orletta, 2020). A core downstream task is dependency parsing: English-specific Transformer pipelines routinely achieve F1 scores above 0.90 (e.g., spaCy's *en_core_web_trf*; Honnibal et al., 2020), and comparable performance has been reported for multilingual Transformer parsers (Van Nguyen et al., 2021). PLMs have also proven effective in SRL. For example, Shi and Lin (2019) demonstrated that a BERT-LSTM model could achieve F1 scores of 0.90 on in-domain test sets and 0.84 on out-of-domain test sets by identifying and classifying arguments without relying on auxiliary syntactic features such as part-of-speech tags or dependency trees.

While dependency parsing and SRL primarily target word-level annotations, PLMs have also shown promise in discourse-level annotation. Eguchi and Kyle (2023), for instance, applied a RoBERTa-based ensemble model to identify and categorize rhetorical stance features in

academic writing. Using a discourse-analytic framework, they manually annotated 4,688 sentences across eight stance categories and trained a model combining RoBERTa and LSTM. The model achieved a macro-averaged F1 score of 0.72 in span identification, outperforming pre-adjudication human annotator agreement.

2.7 Summary

This chapter had laid the conceptual and methodological groundwork for the present study. It first clarified how linguistic complexity is conceptualized in L2 research, distinguishing absolute from relative complexity and differentiating complexity from sophistication (Section 2.1). Adopting a usage-based perspective, the discussion then emphasized that language development is driven by the distributional patterns learners encounter and reproduce, with these patterns spanning a continuum from grammar through lexis, thus motivating the need for a multivariate lexicogrammatical analyses (Section 2.2).

Existing research that focused on measuring L2 productive proficiency has pursued two main perspectives. One line quantifies complexity (i.e., the amount and variety of linguistic structure) through four families of indices: length-based syntactic measures, fine-grained syntactic measures targeting specific clause or phrase relations, constructional diversity measures that count distinct ASCs, and lexical diversity measures (Section 2.3). The second line targets sophistication (i.e., the advancement of forms focusing on frequency and contingency) and clusters usage-based metrics into three groups: construction frequency and verb-construction association strength, n-gram association strength, and multifaceted indices of lexical sophistication (Section 2.4). Complementing both lines, a growing body of work applies ASC/VAC metrics to examine how task design, genre, proficiency, and learning environment

shape learner output, demonstrating that usage-based constructional measures sensitively track developmental change (Section 2.5).

Despite advances in measuring lexicogrammatical complexity and sophistication, (through the development of operationalized indices, NLP tools for automatic calculation, and validation studies), a comprehensive framework for ASC-based measures remains elusive. One key challenge is the reliable and automatic extraction of ASCs and their associated verbs from corpora. Traditional manual annotation is labor-intensive and limited in replicability and scalability (Section 2.6.1). More recent computational approaches have leveraged syntactic and semantic frames as proxies to automate this process. However, these proxies are not always interchangeable; they do not consistently align in a way that transparently captures ASCs (Section 2.6.2). To address this gap, recent advancements in PLMs offer a promising strategy. By applying supervised learning to annotated datasets, PLMs can extract ASCs with enhanced robustness and replicability (Section 2.6.3). The next section (Section 3) introduces a PLM-based ASC tagger, along with a set of operationalized indices to capture ASC usage across learner and reference corpora.

3 Methodology

Portions of this chapter are based on published co-authored work. The annotation work was published in the *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)* in March 2024 as *Annotation Scheme for English Argument Structure Constructions Treebank*, and the project related to ASC tagger was published in the *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* in November 2024 as *Leveraging Pre-trained Language Models for Linguistic Analysis: A Case of Argument Structure Constructions*. I was responsible for identifying the research problem, designing the annotation scheme, constructing the annotated dataset, developing the ASC tagger, and implementing the ASC analyzer. Dr. Kris Kyle provided guidance on model evaluation and interpretation of the findings.

3.1 Tool design

Building on recent advances in leveraging pre-trained language models for linguistic analysis, this section introduces two core components developed through an NLP pipeline: a supervised ASC tagger and an ASC analyzer. The ASC tagger is designed to automatically identify and classify targeted ASC types. It was trained on a gold-annotated dataset and evaluated for robustness across three language-use domains: L1, L2 written, and L2 spoken. In turn, the ASC analyzer takes the tagged output and computes a set of operationalized indices that quantify ASC usage, such as diversity, frequency, and verb-construction associations. Figure 1 provides an overview of the development process for both components.

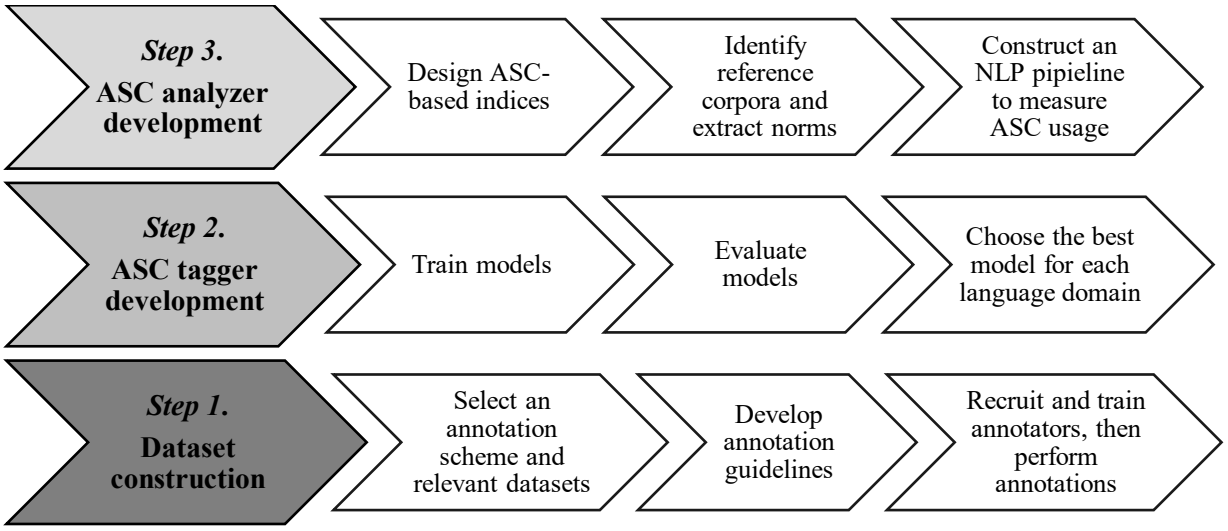

Figure 1. Overview of the development of ASC analyzer

3.2 Dataset construction

3.2.1 Annotation scheme

When developing an annotation scheme, it is important to address two levels of linguistic annotation (cf. Nivre et al., 2004). The first level involves selecting an annotation unit, and the second involves defining the linguistic categories to be used. For this study, each verb-anchored clause served as the unit of annotation, because the verb typically forms the core around which the argument structure is organized (Fillmore, 1968; Goldberg, 1995; see also a similar approach to VAC extraction: Kyle, 2016). In addition, while early theoretical studies on ASCs primarily focused on finite clause types, this study includes both finite clauses and non-finite clauses to capture all ASC-constrained meanings. Consequently, some sentences were parsed into multiple layers—each layer representing a distinct ASC, whether finite or non-finite (Figure 2). Additionally, word order was considered a separate construction layer, which means that different word orders for the same ASC type (for example, due to pragmatic reasons such as

posing a question) were annotated identically.[4] This approach aligns with the understanding that ASCs do not dictate specific word orders (cf. Goldberg, 2019, p. 39).

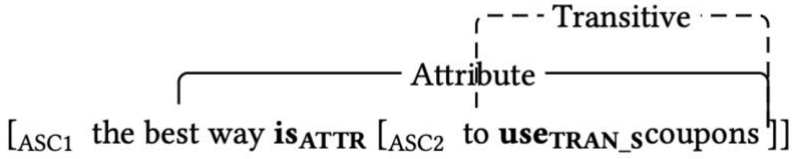

Figure 2. Example of the annotation unit.
Notes. In the visual representation, a solid line denotes an ASC span in a finite clause, while a dashed line indicates an ASC span in a non-finite clause.

For the categories, this study selected nine ASC types. Although previous researchers have varied both the number of ASC types and the complexity of their representations (e.g., Bonial's 2014 focus on light-verb constructions), the main goal of this study was to develop a set of categories broad enough to capture the diverse ASC forms and semantic nuances yet streamlined for effective machine learning. Seven primary ASC types (characterized by the active voice) were considered first, as often outlined in earlier studies (e.g., Bencini and Goldberg, 2000; Ellis and Ferreira-Junior, 2009b; Goldberg, 1995; Hwang and Kim, 2023): (1) intransitive simple (e.g., "worked a lot")[5]; (2) intransitive motion (e.g., "went out the door"); (3) intransitive resultative (e.g., "got healthier"); (4) transitive simple (e.g., "met my friends"); (5)

[4] For example, consider these two expressions with different syntactic frames: (1) "they wanted the thing" and (2) "the thing which they wanted". Both were tagged as a simple transitive construction (TRAN_S), even though the second example features an inverted object ("the thing") due to its relative clause structure. This decision reflects the fact that, despite the surface differences, the semantic roles remain identical (i.e., the verb's relationship to its object is preserved).

[5] The sentences with middle voice (e.g., "The pot broke on the floor") were tagged under the (1) intransitive simple category. While the categorization aligned with the goal of this study, future research may adopt finer-grained approaches to sub-categorize each voice tag. Different voices such as active, middle, passive have the potential to be considered distinct layers (Goldberg, 2013, p. 455). These voices may play an important role in shaping syntactic/semantic frames, altering dynamics of agency and focus within sentences. For example, when a sentence "A car hit the man" is recast in the passive as "The man was hit by a car," the arguments and their semantic roles remain consistent (car-*agent*; man-*theme*), but only the focus within the statement is shifted. This allows for ASCs from active voice categories, which have *agent* and *theme* arguments (i.e., transitive simple, ditransitive, caused-motion, transitive resultative), may have respective passive counterparts and could potentially be grouped into separate passive categories.

ditransitive (e.g., “told them my plan”); (6) caused-motion (e.g., “put the date on the calendar”); (7) transitive resultative (e.g., “made me happy”).[6] Then, two additional types were added to account for ASCs common in L2 learners’ output: (8) attributive (e.g., “be the first”) and (9) passive (e.g., “were recommended by my friend”). Table 3 illustrates these nine ASC types, along with their most prototypical syntactic and semantic frame representations.

Table 3. Target ASCs and semantic-syntactic representations

ASC (TAG)	**Semantic frame**	**Syntactic frame**
Attributive (ATTR)	*theme*-VERB-*attribute*	*nsubj*-*cop*-root
Caused-motion (CAUS_MOT)	*agent*-VERB-*theme*-*destination*	*nsubj*-root-*obj*-*obl*
Ditransitive (DITRAN)	*agent*-VERB-*recipient*-*theme*	*nsubj*-root-*iobj*-*obj*
Intransitive motion (INTRAN_MOT)	*theme*-VERB-*goal*	*nsubj*-root-*obl*
Intransitive simple (INTRAN_S)	*agent*-VERB	*nsubj*-root
Intransitive resultative (INTRAN_RES)	*theme*-VERB-*result*	*nsubj*-root-*advmod*
Passive (PASSIVE)	*theme*-*aux*-$V_{passive}$	*nsubj:pass*-*aux:pass*-root
Transitive simple (TRAN_S)	*agent*-VERB-*theme*	*nsubj*-root-*obj*
Transitive resultative (TRAN_RES)	*agent*-VERB-*theme*-*result*	*nsubj*-root-*obj*-*xcomp*

[6] In this framework, experiential and possessive acts (e.g., X fears Y, X has Y) are treated as subtypes of transitive events, and communicative acts with a receiver (marked by a prepositional phrase) are classified as a form of caused motion, interpreted as the transfer of information from X to Y. Although alternative fine-grained categorizations might differentiate these event types further, the categorization in this study captures the essential argument patterns by which actions and states are expressed in language.

3.2.2 Annotated datasets

Publicly available L1 and L2 English datasets from the Universal Dependency (UD) project served as the basis for ASC annotation.[7] Both datasets utilize the CoNLL-U format,[8] a widely adopted tabular representation for UD annotations. In this format, each sentence is displayed vertically, with one line per token and its linguistic attributes. The format comprises ten columns capturing diverse linguistic information: word index, word form, lemma, universal and language-specific part-of-speech tags, morphological features, syntactic head, dependency relation, enhanced dependency graph, and a miscellaneous column for additional annotations. In the present analysis, the miscellaneous column was used to annotate ASC tags (Figure 3).

```
# text = You called me!
1	You	you	PRON	PRP	Case=Nom	2	nsubj	_	_
2	called	call	VERB	VBD	Tense=Past	2	ROOT	_	TRAN_S
3	me	I	PRON	PRP	Case=Acc	2	dobj	_	_
4	!	!	PUNCT	.	Punct	2	punct	_	_
```

Figure 3. An example of ASC annotation in the CoNLL-U format

For the L1 dataset, English portion of the UP treebank (Akbik et al., 2015) was employed. This combines the UD version of the English Web Treebank (EWT; Silveira et al., 2014) with semantic role labels based on the PropBank annotation scheme. The original EWT corpus comprises sentences sampled from five web registers (blogs, newsgroups, emails, reviews, and *Yahoo!* Answers). A subset of 5,936 sentences (104,640 words) was extracted from the total 16,621 and manually annotated for ASC tags.

The L2 written dataset (Berzak et al., 2016) draws on the CLC FCE dataset (Yannakoudakis et al., 2011), which contains written responses from Cambridge English exams

[7] https://universaldependencies.org/en/
[8] https://universaldependencies.org/format.html

across five registers: letter, report, article, composition, and short story. These samples represent upper-intermediate learners from 10 different native-language backgrounds. From the original 5,124 sentences, 1,948 sentences (37,055 words) were selected and manually tagged.

The L2 spoken dataset (Kyle et al., 2022) was sourced from the NICT JLE corpus (Izumi et al., 2004), which consists of transcriptions of oral proficiency interviews by Japanese English learners. All 2,320 sentences (21,312 words) that included syntactic dependency annotations were extracted and manually tagged.

3.2.3 Annotation guidelines

Because trained human annotators were expected to read the extracted sentences, categorize each sentence into one of the nine ASC types, and annotate their decisions, it was important to develop an informed and consistent annotation scheme, which is a practice commonly found in annotation projects (cf. Gerdes and Kahane, 2016). Accordingly, annotation guidelines were written in advance to provide clear instructions for the recruited annotators. Each ASC category was introduced with a general description of its construction, followed by its syntactic frame depicted through dependency relations based on the UD framework (de Marneffe et al., 2021; Nivre et al., 2016, 2020). In addition, semantic frames derived from the PropBank clarify the roles and relationships of core arguments (cf. Kyle and Sung, 2023, pp. 53-54).

For example, the caused-motion construction (tagged as CAUS_MOT) covers instances in which an agent directly or indirectly induces a theme to move or change location, often via a directional phrase. The guidelines indicate that this construction may involve two types of causation: (1) direct causation: The agent's action immediately instigates the theme's motion (e.g., "She three the ball into the room"), where the act of throwing directly causes the ball's

movement; and (2) indirect causation: The agent's action triggers a series of events or conditions that eventually lead to the theme's movement (e.g., "She scared the cat into the house"), where the act of scaring results in the cat's relocation.

Guidelines were iteratively updated throughout annotation to record rationale, address ambiguous or complex verb-ASC combinations (particularly in L2 data), and supply extensive examples from both L1 and L2 corpora. These updates ensured transparency and consistency.

3.2.4 Annotation process and review

The annotation project spanned eight months (from April to November 2023). Six undergraduate linguistics majors (native English speakers who had completed advanced coursework in functional English syntax) were recruited and trained over three structured, one-hour sessions. In the first session, annotators received an overview of the project's theoretical background, objectives, and their roles, along with the annotation guidelines. They were also introduced to the CoNLL-U data format, the tagging scheme, and the procedures for accessing and storing files in a shared folder. Standard text editors (e.g., *BBEdit*) and spreadsheet software (e.g., *Microsoft Excel*) provided a straightforward interface for manual tagging.

During the subsequent sessions, annotators practiced on sample sentences, tagging each item individually and then discussing challenges as a group. After each exercise, the researcher provided feedback. Once training was complete, annotators worked remotely, uploading their files to a monitored shared folder. To ensure consistency and address questions, the guidelines were updated weekly based on annotator performance, and a dedicated *Discord* server was maintained for real-time support.

Although the original UD annotations (e.g., *nsubj*, *root*) offered useful syntactic cues, annotators treated them as flexible guidelines—mapping UD tags to ASC categories when reliable, but overriding them when the syntactic structure did not align with the construction's semantic arguments. For instance, in the case of *enter*-type verbs, even when the UD structure may suggest a simple transitive frame (i.e., *nsubj*-root-*dobj*), the semantic roles conveyed often point to a different constructional interpretation of intransitive motion construction, as they typically present a theme in the subject position followed by a destination (e.g., "enter the classroom").

Repetitive borderline cases were fully documented and interpreted in the ASC annotation manual (Chapter 5, p. 24), which lists each ambiguous construction and the relation for its final annotation. Each token was annotated independently by two randomly assigned annotators in a blind review. Disagreements triggered a third review by another annotator or the researcher; if disagreement persisted, the researcher performed a fourth adjudication.

The inter-annotator agreement during the first round of annotation was reasonably strong: exact agreement was 85.7%, with Cohen's kappa=.801 (Landis and Koch, 1977).[9] Table 4 reports the number of annotated ASCs in each dataset, and Figure 4 shows a confusion matrix for annotator agreement by ASC tag.

Table 4. ASCs distribution in the gold-standard ASC treebank

TAG	L1	L2 written	L2 spoken
ATTR	2,539	1,289	760
CAUS_MOT	766	87	53
DITRAN	285	160	37
INTRAN_MOT	607	250	240
INTRAN_S	1,395	662	525
INTRAN_RES	213	44	23
PASSIVE	1,058	224	50

[9] These figures were slightly lower when misspelled and missed tags were included (exact agreement = 82.5%, kappa = .759). These are not represented in the confusion matrix in Figure 4.

TRAN_S	6,094	2,488	1,385
TRAN_RES	763	76	16
Total	**13,720**	**5,260**	**3,089**

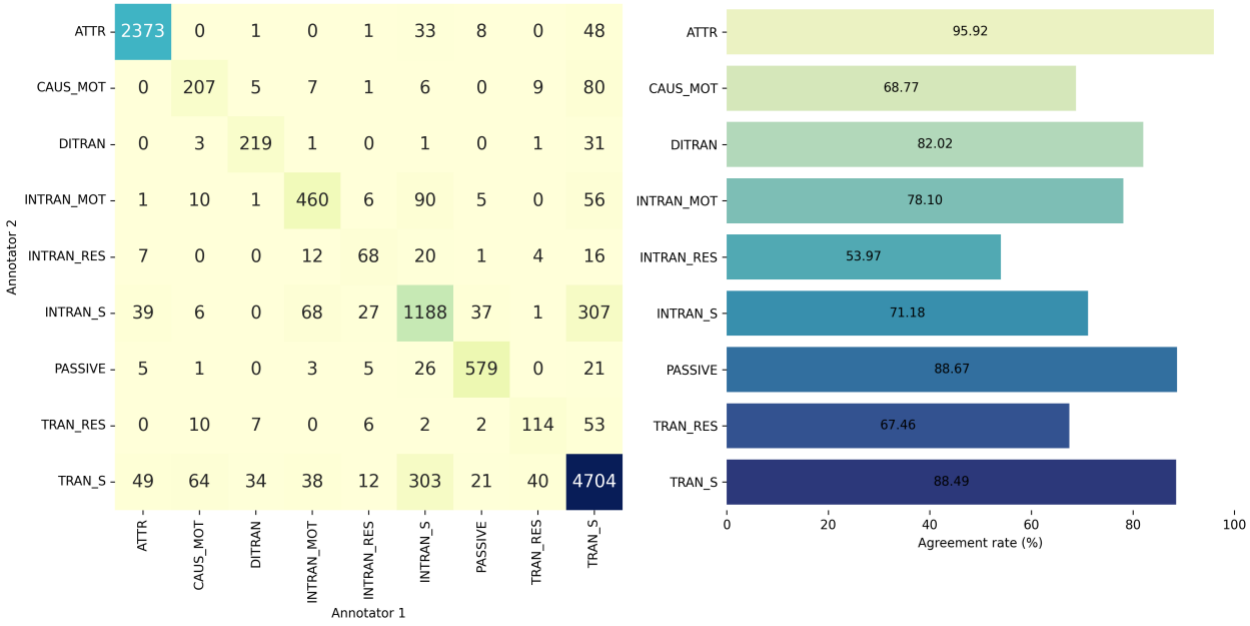

Figure 4. Evaluation of inter-annotator agreement

3.3 ASC tagger development

3.3.1 Training setup

The gold-annotated ASC treebank (constructed as described in Section 3.2) was used to train a multi-class named entity recognition (NER) model with spaCy (version 3.7.4; Honnibal et al., 2020). The model builds on RoBERTa embeddings via the en_core_web_trf pipeline, which integrates pre-trained transformer weights and fine-tunes them on the ASC annotations. In addition, spaCy's transition-based parser provides syntactic and semantic context by modeling linguistic structures through a learned sequence of deterministic transitions, further improving NER performance. All models shared identical hyperparameter settings; see Appendix A for details.

3.3.2 Model evaluation setup

In addition to evaluating the ASC tagger's accuracy on the test set, its robustness was assessed by comparing it with a prior semantic role-based extraction method (described in Section 2.3.2.2). That earlier approach outperformed earlier syntactic-frame (and syntactic-frame+lemma-based) methods in ASC extraction (Kyle & Sung, 2023). To facilitate this comparison, two versions of the ASC treebank were compiled: a *silver*-annotation based on the dataset provided by Kyle & Sung (2023), and a *gold*-annotation built from this gold-annotation project. The silver version comprises 26,437 ASC tokens that were semi-automatically annotated using semantic role labeling, while the gold version consists of 22,069 tokens that were manually annotated. Although the two treebank versions differ methodologically, they are similar in size, enabling a reasonably fair comparison of training and evaluation performance.

For training, the silver-annotated dataset was partitioned into 80% training, 10% development, and 10% test sets. To examine how language-use domain diversity affects model performance, the gold annotations were further divided into two subsets: (1) L1 only, and (2) combined L1+L2 (written and spoken registers). Because L2 data were underrepresented, the L1 + L2 subset was resampled to 34 % training, 33 % development, and 33 % test, whereas the L1-only subset retained the original 80/10/10 distribution. Thus, the silver model used an 80/10/10 split on silver data; the gold L1 model used an 80/10/10 split on L1 data; and the gold L1+L2 model used a 34/33/33 split on combined data—adding 2,835 training tokens, 2,765 development tokens, and 2,758 test tokens relative to the gold L1 model. Table 5 shows the ASC distributions in the gold datasets by domain.

Table 5. Distribution of ASCs in the gold-standard treebank across different domains

ASC	L1	L2 written	L2 spoken

	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
ATTR	2,058	258	223	399	445	445	242	266	252
CAUS_MOT	641	61	64	26	30	31	18	18	17
DITRAN	235	31	19	59	47	54	16	12	9
INTRAN_MOT	502	55	50	86	79	85	91	82	67
INTRAN_RES	172	23	18	15	13	16	11	5	7
INTRAN_S	1,154	135	106	243	209	210	146	190	189
PASSIVE	867	102	89	72	59	73	16	18	16
TRAN_RES	622	77	64	25	23	23	6	5	5
TRAN_S	4,900	598	596	858	824	806	506	426	453
Total	11,151	1,340	1,229	1,783	1,734	1,743	1,052	1.022	1,015

3.3.3 Model evaluation result

Table 6 demonstrates that the gold L1+L2 model achieved the highest weighted average F1 scores across all test sets—L1 (F1 = 0.912), L2 written (F1 = 0.915), and L2 spoken (F1 = 0.928)—and attained the top F1 for seven of nine ASC tags in both L2 registers. These findings confirm that models trained on manually annotated gold-standard data consistently outperform those trained on silver-standard data. Consequently, the gold L1+L2 model was chosen for L2 tagging, owing to its superior performance, while the gold L1 model was selected for L1 tagging, as it delivers a more balanced tag accuracy despite slightly lower scores for CAUS_MOT, INTRAN_S, PASSIVE, and TRAN_S.

Table 6. F1 scores across ASC types, models, and domains

ASC	silver			gold L1			gold L1+L2		
	L1	L2 w	L2 s	L1	L2 w	L2 s	L1	L2 w	L2 s
ATTR	**0.982**	0.955	0.971	0.972	0.954	0.986	0.968	**0.971**	**0.988**
CAUS_MOT	0.794	0.764	0.690	0.818	0.833	**0.710**	**0.857**	**0.867**	**0.710**
DITRAN	0.757	0.862	**1.000**	**0.919**	**0.914**	0.842	0.865	0.881	0.947
INTRAN_MOT	0.763	0.755	0.774	**0.800**	0.770	0.789	0.772	0.807	**0.843**
INTRAN_RES	0.667	0.741	0.000	**0.750**	0.788	0.800	0.625	0.813	**0.833**
INTRAN_S	0.806	0.770	0.853	0.779	**0.806**	0.817	**0.808**	0.803	**0.865**
PASSIVE	0.932	**0.865**	0.875	0.920	0.775	**0.938**	**0.940**	0.865	0.909
TRAN_RES	0.853	0.714	0.588	**0.884**	0.800	**0.625**	0.881	0.792	**0.625**
TRAN_S	0.922	0.904	0.933	0.931	0.929	0.927	**0.936**	0.943	**0.948**
Weighted Average	0.902	0.885	0.907	0.908	0.900	0.905	**0.912**	**0.915**	**0.928**

Notes. L2 w(ritten); L2 s(poken); The highest scores per ASC tag in each dataset (L1, L2w, L2s) are bolded.

3.4 ASC analyzer development

With the ASC tagger validated, the subsequent step involved usage analysis. An ASC analyzer was constructed using four types of indices (i.e., diversity, proportion, frequency, and SOA) to capture both internal patterns within learner texts and alignment with reference-corpus norms (Figure 5). Diversity and proportion serve as text-internal measures, reflecting the variety and distribution of ASCs within each text. Frequency and SOA function as text-external measures, comparing observed ASC usage against advanced language benchmarks from a reference corpus.

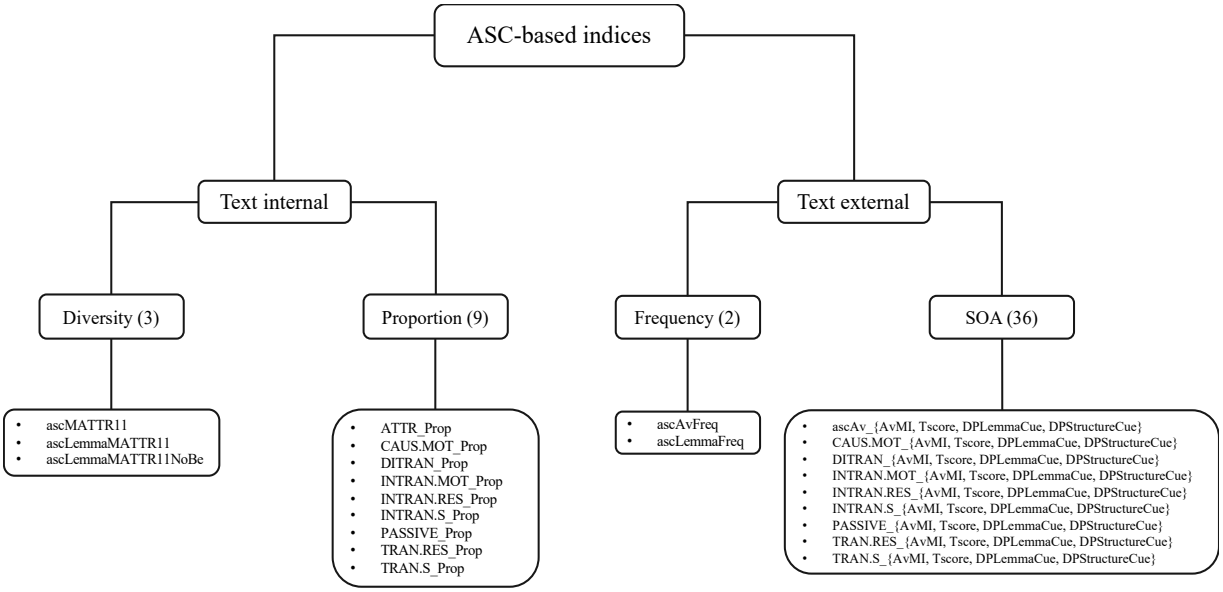

Figure 5. Categorizations of ASC-based indices

A description of each index is provided below. Based on these descriptions, Python code was written to extract the target features.

Diversity

- *ascMATTR11* measures the moving average type-token ratio over an optimized window of 11 ASCs, assessing the ASC diversity within the text.
- *ascLemmaMATTR11* measures moving average type-token ratio over an optimized window of 11 ASCs, assessing the ASC-verb lemma combination diversity within the text.
- *ascLemmaMATTR11NoBe* excludes forms of the verb "be" from the ascLemmaMATTR11 diversity calculation. Since "be" appears frequently in corpora but contains less semantic information compared to other verbs, this index provides a picture of the diversity of more semantically rich verbs and content words.

Proportion

- *ATTR_Prop* measures the proportion of attributive constructions within the text.
- *CAUS.MOT_Prop* measures the proportion of caused-motion constructions.
- *DITRAN_Prop* measures the proportion of ditransitive constructions.
- *INTRAN.MOT_Prop* measures the proportion of intransitive motion constructions.
- *INTRAN.RES_Prop* measures the proportion of intransitive resultative constructions.
- *INTRAN.S_Prop* measures the proportion of simple intransitive constructions.
- *PASSIVE_Prop* measures the proportion of passive constructions.
- *TRAN.RES_Prop* measures the proportion of transitive resultative constructions.
- *TRAN.S_Prop* measures the proportion of simple transitive constructions.

Frequency

- *ascAvFreq* measures the frequency of ASCs in the text, based on the averaged frequency calculated from the reference corpus.

- *ascLemmaFreq* measures the frequency of ASCs-verb lemmas in the text, based on the averaged frequency calculated from the reference corpus.

Strength of association (SOA)

The SOA of ASCs and verb lemma based on the database is calculated from the reference corpus. Three types of indices are included: mutual information (MI), t-score, and Delta P. These were operationalized based on previous studies that measured SOA in n-grams (e.g., Eguchi & Kyle, 2020; Kyle & Eguchi, 2021), dependency bigrams (e.g., Kyle & Eguchi, 2023), or syntactic sophistication (e.g., Kyle, 2016; Kyle & Crossley, 2017). All the values are averaged (arithmetic mean) across the ASC-verb lemma pairs to yield an averaged scores for each construction (e.g., CAUS.MOT_AvMI) and for all ASC pairs (e.g., ascAvMI). The indices and their formulas are as follows:

- Pointwise mutual information (*AvMI*): $log_2\left(\frac{observed}{expected}\right)$

  AvMI measures, on average, how much more information an ASC-verb lemma pair is than would be expected by chance. First, for each ASC and verb lemma, we count how often they co-occur in the reference corpus (observed frequency). Next, we estimate how often they would co-occur if their occurrences were independent (expected frequency). We then convert the ratio of observed to expected frequency into bits of information by using a base-2 logarithm. Finally, we average these bit-values across all ASC-verb lemma pairs in the learner text to produce the score.

- T-score (*Tscore*): $\frac{observed-expected}{\sqrt{observed}}$

  Tscore is computed in a manner analogous to AvMI, using both observed and expected co-occurrence frequencies. For each ASC-verb lemma pair, we take its actual count in the reference corpus and subtract the count predicted under statistical independence. We then

divide this difference by the square root of the observed frequency. This normalization down-weights low-frequency pairs and yields a standardized association score: higher T-scores indicate that the ASC-verb combination co-occurs more reliably than chance would predict.

- Delta P measures the directional predictability of one element by another, given then 2x2 table to describe the equation:

	Target ASC present	Target ASC absent
Target verb lemma present	*a*	*b*
Target verb lemma absent	*c*	*d*

  - Delta P Lemma Cue (*DeltaLemmaCue*): $\frac{a}{a+b} - \frac{c}{c+d}$

    DeltaLemmaCue measures the change in the probability of the target ASC occurring when the target lemma is present, compared to the probability of the target ASC occurring when the lemma is absent.

  - Delta P Structure Cue (*DPStructureCue*): $\frac{a}{a+c} - \frac{b}{b+d}$

    DPStructureCue measures the change in the probability of the target lemma occurring when the target ASC is present, compared to the probability of the target lemma occurring when the ASC is absent.

- *ascAv_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the average scores of various SOA metrics for all ASC-lemmas in the text, with the scores being calculated for each ASC-lemma combination and then macro averaged.
- *CAUS.MOT_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for caused-motion constructions.
- *DITRAN_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for ditransitive constructions.

- *INTRAN.MOT_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for intransitive motion constructions.
- *INTRAN.RES_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for intransitive resultative constructions.
- *INTRAN.S_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for simple intransitive constructions.
- *PASSIVE_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for passive constructions.
- *TRAN.RES_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for transitive resultative constructions.
- *TRAN.S_{AvMI, Tscore, DPLemmaCue, DPStructureCue}* measures the scores for simple transitive constructions.

3.5 Summary

This chapter outlined the methodology behind the ASC analyzer. It opened with an overview of the tool's architecture and its core component, the ASC tagger, which automates the identification and classification of nine ASC types. The overview also outlined the tool's developmental workflow and its evaluation framework across three language-use domains including L1, L2 written, and L2 spoken (Section 3.1).

Next, the construction of the annotated dataset that underpins the ASC tagger was detailed (Section 3.2). The annotation scheme, the creation of comprehensive annotation guidelines, and the iterative processes of dataset curation were explained. These steps ensured that the training data are both robust and representative of the targeted ASC types.

The focus shifted to the development and training of the ASC tagger (Section 3.3). The experimental setup, model training procedures, and evaluation metrics were presented, and the resulting performance metrics underscored the tagger's effectiveness while highlighting areas for future refinement.

Finally, the ASC analyzer was introduced as an extension of the tagger (Section 3.4). Four types of the ASC index were defined and employed to analyze ASC usage. With this methodology in place, the study proceeds to two empirical investigations: one based on L2 speaking assessment (Section 4) and the other on L2 writing assessment (Section 5).

4 Measuring lexicogrammatical complexity and sophistication in L2 speaking assessment

This chapter is based on co-authored work that has been accepted for publication in *Studies in Second Language Acquisition* as *Usage-based Analysis of L2 Oral Proficiency: Characteristics of Argument Structure Construction Use* (in press). I was responsible for identifying the research questions, conducting the data analysis, and interpreting the results. Dr. Kris Kyle provided feedback on the analytical approach and interpretation.

Indices of lexicogrammatical complexity and sophistication, including measures at the lexical and n-gram levels, have long played a crucial role in assessing L2 oral proficiency. Several studies have focused on lexical measures of diversity (Kyle et al., 2024) and sophistication (Eguchi & Kyle, 2021), identifying indices such as MATTR and word-frequency, which together enhance the multivariate explanation of L2 oral proficiency. Beyond the word level, research on n-grams, specifically their frequency and association strength, has provided additional insights into proficiency variation (Eguchi & Kyle, 2021). Moreover, investigations at the lexicogrammatical interplay employing dependency-based bigram measures have shown that these combined metrics account for a significant proportion of variance in L2 oral proficiency (Kyle & Eguchi, 2023).

At the clausal level, while a few studies have investigated ASCs as potential predictors of fluency (Choi & Sung, 2020) and advanced speaking proficiency (Kim & Ro, 2023), most research has focused narrowly on the structural complexity of grammatical constructions produced by L2 speakers. Methodological challenges in ASC extraction have also constrained their broader application, and the predictive validity of ASC-based indices remains underexplored in multivariate frameworks. To address these research gaps, this chapter is guided by the following RQs:

RQ 1. What is the relationship between L2 oral proficiency scores and ASC-based indices?

RQ 2. To what extent are ASC-based indices predictive of L2 oral proficiency scores?

RQ 3. To what extent are ASC-based indices predictive of L2 oral proficiency scores when combined with other lexicogrammatical indices?

4.1 Method

4.1.1 Datasets

4.1.1.1 Learner corpus

The National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus was used for this study. The corpus comprises 1,281 oral proficiency interview transcripts from Japanese learners of English (Izumi et al., 2004). The interviewees participated in the Standard Speaking Test (SST), which was modeled after the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL-OPI) and subsequently adapted for Japanese learners of English (ACTFL-ALC Press, 1996; Koizumi & Hirai, 2012, pp. 41-42). The SST comprises five sections: a self-introduction, a description of a picture, a role-play, a narration based on a series of images, and a concluding section with reflective questions. Each response was evaluated by at least two qualified raters, with a third rater resolving any score discrepancies (Kobayashi & Abe, 2016, p. 59). Test-takers received a holistic score ranging from 1 to 9, along with feedback on various criteria, including task completion, interaction quality, text type, grammatical accuracy, pronunciation, and interviewers' comments. A comprehensive scoring rubric is provided in Appendix B.[10]

[10] For further description of the dataset, see also https://alaginrc.nict.go.jp/nict_jle/index_E.html.

The original dataset includes tags pertaining to various aspects of the production, such as the interview structure (e.g., different stages of the interview), metadata (e.g., age, occupation, proficiency level), speaker turns (indicating whether utterances were made by interviewers or interviewees), discourse phenomena (e.g., repetition, self-correction), and grammatical errors (e.g., verb tense errors). Based on these tags, additional filtering was performed to remove linguistic features that were not pertinent to the current analysis, such as fillers, repetitions, Japanese words/utterances, and paralinguistic cues. Replicating the approach of Kyle et al. (2024, p. 7), publicly available Python code from that study was utilized to extract a clean text version from interview stages 2 to 5, which include the single-picture description, role-play, and sequential picture storytelling tasks.

Table 7 presents the distribution of learners and token statistics across different proficiency score levels. The data include mean, minimum, maximum, and standard deviation of the tokens. Proficiency scores ranged from 1 to 9, with a mean of 4.664, a standard deviation of 1.574, a median of 4, and a standard error of 0.044.

Table 7. Number of learners and tokens across L2 oral proficiency levels

Proficiency level	**Number of learners**	**Tokens (mean)**	**Tokens (min)**	**Tokens (max)**	**Tokens (std)**
1	3	81.67	60	111	30.07
2	35	125.54	51	240	43.90
3	222	279.04	101	533	80.16
4	482	428.03	219	798	101.99
5	236	584.32	310	1041	132.55
6	130	688.23	383	1094	148.82
7	77	726.13	510	1238	150.55
8	56	851.54	555	1401	204.72
9	40	964.90	576	1650	228.76

4.1.1.2 Reference corpus

The *US English subtitle corpus* (SUBTLEX-US) was used as a reference to compute frequency and association strength norms. These norms were employed to determine how common or rare verb lemmas, ASCs, and their strength of associations are in the learner corpus compared to those of advanced English speakers. The original corpus was compiled from the subtitles of films and television programs, totaling 8,388 subtitle files (Brysbaert & New, 2009, Brysbaert et al., 2012), and contains 76,965,430-word tokens, 164,686-word types, 5,128,462 sentences, and 5,665,251 tagged ASCs. Although subtitles are scripted and edited for readability, they nonetheless approximate spontaneous spoken language across diverse narrative contexts and have been used as a benchmark in L2 oral studies (cf. Berger et al., 2019, p. 919, "derived from a corpus of American film and television subtitles approximating spoken language but excluding highly formal or academic written registers"). Figure 6 shows the normalized frequencies of different ASCs in the learner and the reference corpora, highlighting distributional differences across ASC types.

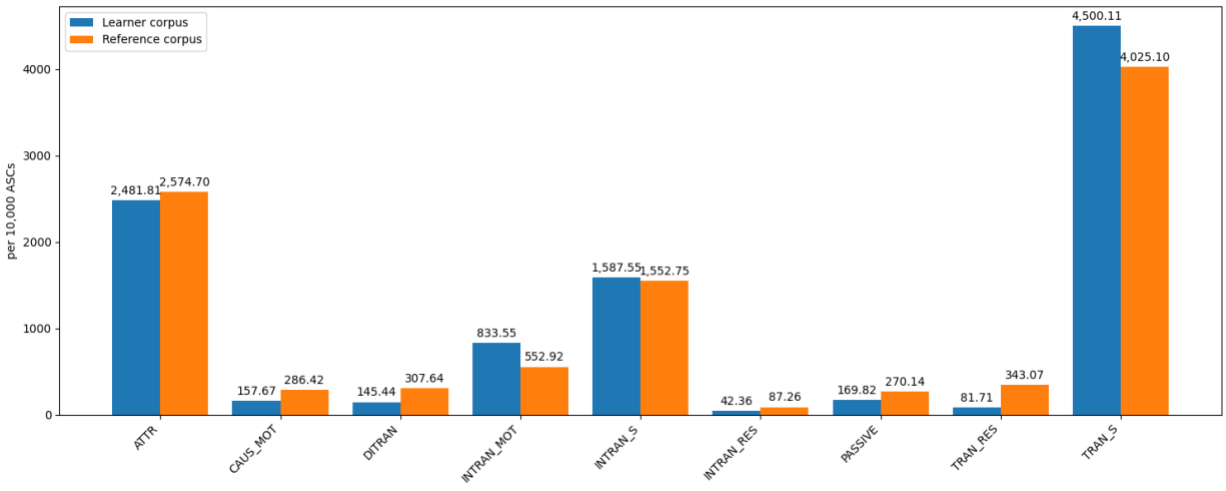

Figure 6. Distribution of ASCs in the JLE (learner) and SUBTLEX-US (reference) corpora
Notes. Frequency counts normalized by occurrences per 10,000 ASCs

4.1.2 Target constructs and indices

To quantify lexicogrammatical complexity and sophistication at the ASC-verb interplay, the ASC analyzer computes four types of indices: diversity, proportion, frequency, and SOA (see Section 3.4 for full descriptions). In addition to these ASC-based indices, established lexicogrammatical indices shown to predict L2 proficiency (Eguchi & Kyle, 2020; Kyle & Eguchi, 2023; Kyle et al., 2024) were included. These supplementary indices span three sub-constructs: lexicogrammatical sophistication (e.g., dependency bigram SOA), lexical sophistication (e.g., word frequency, bigram SOA), and lexical diversity (e.g., MATTR). Below, detailed descriptions of each index within its respective sub-construct are provided.

4.1.2.1 Lexicogrammatical sophistication

To measure lexicogrammatical sophistication, the SOA scores for dependency bigrams were computed, focusing on four key bigram dependency structures (Kyle & Eguchi, 2023): *noun-adjective* (*n_amod*), VERB-*adverb* (*v_advmod*), VERB-*direct_object* (*v_dobj*), and VERB-*subject* (*v_nsubj*). These scores were calculated using lemmatized words, representing the average SOA scores for all dependency bigrams in a learner's utterance that appeared at least five times in the spoken COCA, a benchmark reference corpus (Davies, 2010). For each dependency type, the SOA scores were calculated using four SOA indices: T score, MI score, MI^2 score, delta P (calculated based on the dependency relationships of head as a cue (*deltap_*govecue) or dependent as a cue (*deltap_*depcue) or selecting the maximum value between the two (*deltap_strgst*).

- *n_amod_*{T, MI, *MI^2, deltap_govcue, deltap_depcue, deltap_strgst*}
- *v_advmod_*{*T, MI, MI^2, deltap_govcue, deltap_depcue, deltap_strgst*}
- *v_dobj_*{*T, MI, MI^2, deltap_govcue, deltap_depcue, deltap_strgst*}

- *v_nsubj_{T, MI, MI^2, deltap_govcue, deltap_depcue, deltap_strgst}*

4.1.2.2 Lexical sophistication (bigram)

To measure bigram-level lexical sophistication (Eguchi & Kyle, 2020), bigram SOA scores were computed for lemmatized and raw bigrams. Each score reflects the average SOA of all bigrams in a learner's utterance that occur at least five times in the spoken COCA (Davies, 2020). Following the same procedure used for dependency bigram SOA, the bigram SOA scores were derived suing the following indices:

- *lemma_bg_{T, MI, MI^2}* measures lemmatized bigram {T score, MI, MI^2}.
- *lemma_bg_deltap_w1cue* measures delta P, specifically the difference between the conditional probabilities P($word_{right}$|$word_{left}$) and P($word_{right}$|not_$word_{left}$).
- *lemma_bg_delap_w2cue* measures delta P, specifically the difference between the conditional probabilities P($word_{left}$|$word_{right}$) and P($word_{left}$|not_$word_{right}$).
- *lemma_bg_deltap_strgst* measures delta P, calculated in both directions (from $word_{left}$ to $word_{right}$ and from $word_{right}$ to $word_{left}$), selecting the maximum value.

4.1.2.3 Lexical sophistication (word)

Lexical sophistication was also measured at the word level using three indices drawn from prior studies (Eguchi & Kyle, 2020; Kyle & Eguchi, 2023). First, word *frequency* indices were logarithmically transformed to accommodate the Zipfian distribution of the reference corpus. Frequencies were computed for all content words and specific parts of speech, including adjectives (*amod*), adverbs (*advmod, adv_manner*), nouns (*noun, cw_lemma*), and lexical verbs (*mverb*, *lex_mverb*).

- *amod_freq_log* measures the logarithmic frequency of adjectives.
- *advmod_freq_log* measures the logarithmic frequency of adverbs.

- *adv_manner_freq_log* measures the logarithmic frequency of manner adverbs.
- *mverb_freq_log* measures the logarithmic frequency of main verbs.
- *lex_mverb_freq_log* measures the logarithmic frequency of lexical main verbs.[11]
- *noun_freq_log* measures the logarithmic frequency of nouns.
- *cw_lemma_freq_log* measures the logarithmic frequency of all content words.

Word *concreteness* was calculated using Brysbaert et al.'s (2014) norms, which provide ratings for approximately 40,000 English words based on surveys of native speakers (roughly 20 ratings per word). In these norms, higher concreteness scores (e.g., for "apple" or "cookie") denote greater perceptual salience and, theoretically, easier learnability; lower scores (e.g., for "doubt" or "pride") indicate more abstract concepts. The index *b_concreteness* reflects these norms, with larger values signifying more concrete vocabulary.

Contextual distinctiveness was assessed through two complementary indices. First, *mcd* (McDonald & Shillcock, 2001) quantifies the predictability of a word's surrounding five-word window via relative entropy: higher mcd values signal that a word occurs in more distinctive, less predictable contexts. Second, *usf* (Nelson et al., 2004) derives from behavioral association norms, measuring how frequently a word is produced in free-association tasks; lower usf values correspond to greater contextual distinctiveness, since rare associations imply a more unique contextual profile.

4.1.2.4 Lexical diversity

One robust measure of lexical diversity is the index of lexical variety, which reflects the TTR within a text. To mitigate text-length effects, MATTR with an optimized window length

[11] These verbs include any word tagged as a verb, excluding those with an auxiliary dependency tag (such as the copular "be").

was recommended in prior research (Kyle et al., 2024). Accordingly, an 11-word window MATTR index was employed in this analysis.

4.1.3 Statistical analyses

To address RQ 1, which examines the relationship between each ASC-based index and L2 oral proficiency scores, correlation analyses were conducted. Initially, the distribution of all indices was visually inspected using histograms and scatter plots. Indices exhibiting clear skewness, indicative of potential floor or ceiling effects, were excluded, and Pearson correlations were calculated between the remaining indices and the proficiency scores. All analyses were performed using R (*corrtable* package, van der Laken & Lambert, 2023).

To address RQ 2 and RQ 3, which investigate the predictive power of ASC-based indices (RQ 2) and a combination of ASC-based and other lexicogrammatical indices (RQ 3), linear regression analyses were conducted. The process involved several stages of statistical filtering to identify the optimal subset of indices for model building. First, correlations between the variables and the scores were calculated, and variables with an absolute correlation of $|r| < .10$ (indicating a negligible effect; Cohen, 1988) were excluded. Next, for each SOA index type, only the index with the strongest correlation to L2 oral proficiency was retained. The remaining indices were then examined for multicollinearity by assessing their inter-correlations; when two indices exhibited a correlation > .90, the index with the stronger relationship to proficiency was kept. Multicollinearity was further evaluated using the Variance Inflation Factor (VIF), and indices with VIF > 5 were removed.[12]

[12] This threshold is commonly used to identify moderate multicollinearity and ensure robust regression estimates while minimizing the risk of inflated standard errors (James et al., 2017; Menard, 2001).

Following the identification of the optimal index subsets for RQ 2 and RQ 3, an initial multiple-linear-regression model was built in R with *lm*. To guard against over-parameterization and multicollinearity, a model-selection search was carried out with *dredge* function (*MuMIn* package; Barton, 2023). This function evaluates every possible combination of the candidate predictors and ranks the resulting models by Akaike's Information Criterion (AIC). Because AIC rewards goodness-of-fit while penalizing model complexity (Akaike, 1974), smaller values indicate a model that balances explanatory power and parsimony.

Following the guidelines of Tan and Biswas (2012), only models whose AIC values lay within $\Delta AIC < 4$ of the global minimum were retained. This threshold defines the set of "plausible" models that remain empirically competitive with the top-ranked specification while excluding clearly inferior alternatives. Within the $\Delta AIC < 4$ subset, the model with the absolute minimum AIC was selected.[13] Standard diagnostic checks confirmed that the final model met the assumptions of linear regression: a Q-Q plot of standardized residuals indicated approximate normality, a residuals-versus-fitted plot showed no discernible pattern and thus supported homoscedasticity, and variance-inflation factors (all $VIF < 3$) verified an acceptable level of multicollinearity.

After the final model had been selected, the relative importance of each predictor was assessed with the *lmg* (i.e., Lindeman-Merenda-Gold metric; Lindeman, 1980) implemented in the *relaimpo* package (Grömping, 2023). *lmg* averages the incremental increase in R^2 that a predictor contributes across every one of the $p!$ possible orders in which the p predictors could enter the model; the resulting weights therefore partition the model's total R^2 into non-

[13] If two or more models shared the same minimum AIC, parsimony and theoretical interpretability were used as secondary filters, the model retaining the fewest predictors and/or the indices most strongly motivated by prior research was preferred. All candidate models were retrieved with *get.models*() in MuMIn, their AIC values compared, and the lowest-AIC specification was retained for subsequent diagnostics.

overlapping, non-negative shares. Because these raw *lmg* weights sum exactly to the model R^2 (e.g., 0.4), they were rescaled to 100 % for easier comparison across predictors. In this percentage-of-explained-variance scale, for example, a weight of 20% means that the predictor is responsible for one-fifth of the model's explanatory power; the absolute share of total score variance is obtained by multiplying this percentage by the model R^2 (20% × 0.4 = 0.08, or 8 percentage points).

4.2 Results

4.2.1 RQ 1: Relationship between ASC-based indices and L2 oral proficiency scores

To address RQ 1, descriptive statistics and correlations between oral proficiency scores and each index are presented. The key results for each index group: diversity, proportion, frequency, and SOA were summarized in Figure 7, with detailed results for each index type following in each section.

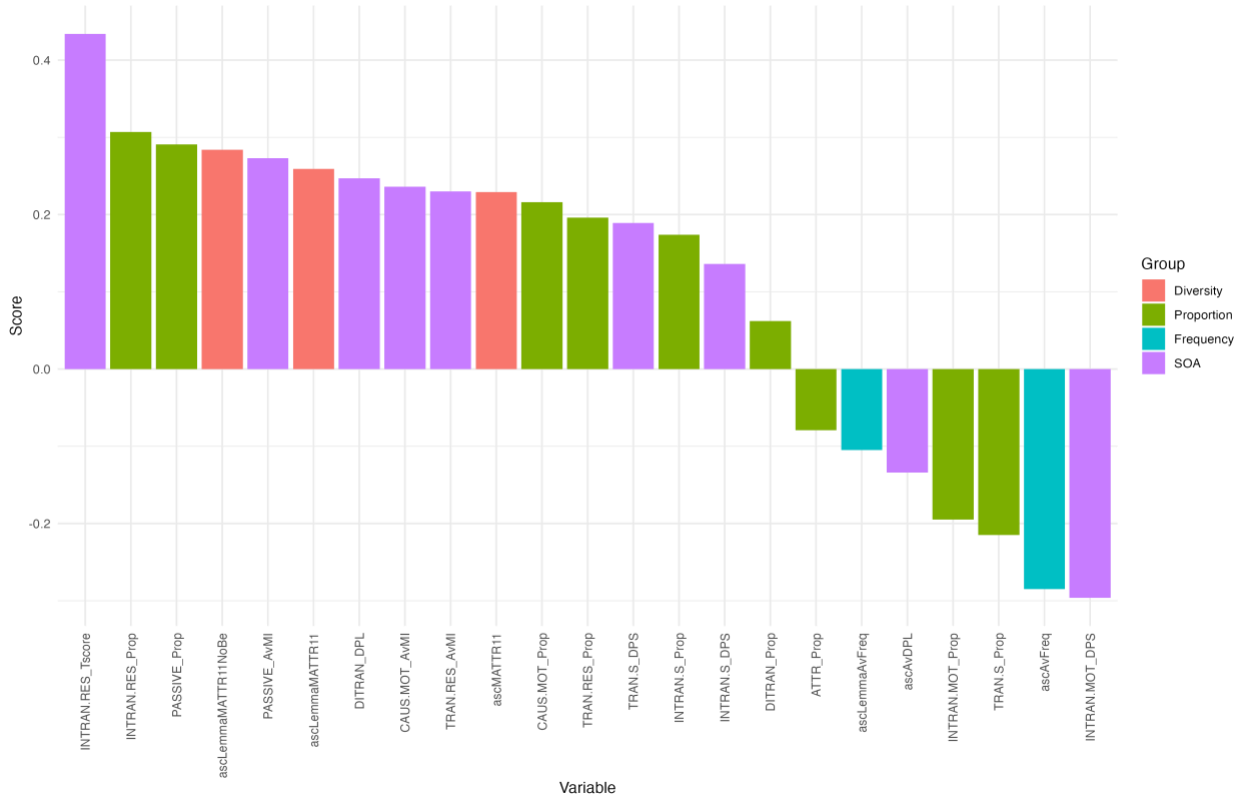

Figure 7. Correlations between L2 oral proficiency scores and selected indices of ASC use

4.2.1.1 Diversity

The descriptive statistics for diversity indices are provided in Table 8. The results of the correlation analysis between the diversity indices and proficiency scores are reported in Table 9. The result indicates that all investigated indices have significant correlations with oral proficiency, but these correlations are generally small (Cohen, 1988). The correlation matrix shows that ascLemmaMATTR11 and ascLemmaMATTRT11NoBe are collinear.

Table 8. Descriptive statistics: Diversity indices

	mean	**min**	**max**	**std**	**SE**
ascMATTR11	0.344	0.244	0.444	0.028	0.001
ascLemmaMATTR11	0.682	0.485	0.909	0.054	0.002
ascLemmaMATTR11NoBe	0.709	0.500	0.889	0.058	0.002

Table 9. Correlations between diversity indices and L2 oral proficiency scores

	Score	ascMATTR11	ascLemmaMATTR11
ascMATTR11	0.229	1.000	
ascLemmaMATTR11	0.259	0.229	1.000
ascLemmaMATTR11NoBe	0.284	0.217	0.864

4.2.1.2 Proportion

The descriptive statistics for proportion indices are presented in Table 10. The results of the correlation analysis between each type of proportions and proficiency scores are detailed in Table 11. The result indicates all investigated indices have significant correlations with oral proficiency and these correlations are generally small to medium. In terms of direction, attribute, intransitive motion and transitive simple ASCs correlated negatively; all others correlated positively.

Table 10. Descriptive statistics: Proportion indices

	mean	min	max	std	SE
ATTR_Prop	0.881	0.000	0.925	0.069	0.002
CAUS.MOT_Prop	0.011	0.000	0.083	0.013	0.000
DITRAN_Prop	0.011	0.000	0.078	0.013	0.000
INTRAN.MOT_Prop	0.070	0.000	0.281	0.038	0.001
INTRAN.RES_Prop	0.003	0.000	0.077	0.006	0.000
INTRAN.S_Prop	0.330	0.148	0.650	0.058	0.002
PASSIVE_Prop	0.012	0.000	0.081	0.014	0.000
TRAN.RES_Prop	0.006	0.000	0.055	0.009	0.000
TRAN.S_Prop	0.362	0.167	0.594	0.066	0.002

Notes. Prop: proportion

Table 11. Correlations between proportion indices and L2 oral proficiency scores

	Score	ATTR	CAUS. MOT	DITRAN	INTRAN. MOT	INTRAN. RES	INTRAN.S	PASSIVE	TRAN. RES
ATTR	-0.079	1.000							
CAUS.MOT	0.216	-0.044	1.000						
DITRAN	0.062	-0.010	0.043	1.000					
INTRAN.MOT	-0.195	-0.036	-0.152	-0.047	1.000				
INTRAN.RES	0.307	-0.077	0.083	-0.014	-0.061	1.000			
INTRAN.S	0.174	-0.063	-0.008	-0.128	-0.175	0.046	1.000		

PASSIVE	0.291	-0.024	0.052	-0.038	-0.118	0.109	0.029	1.000	
TRAN.RES	0.196	-0.042	0.078	0.000	-0.100	0.046	-0.003	0.067	1.000
TRAN.S	-0.215	0.055	-0.096	-0.003	-0.161	-0.114	-0.537	-0.192	-0.074

4.2.1.3 Frequency

The description statistics for the frequency indices are provided in Table 12. The results of the correlation analysis between each type of frequency indices and proficiency scores are detailed in Table 13. The result indicates that all indices correlated significantly with oral proficiency, though the effects were generally small and negative.

Table 12. Descriptive statistics: Frequency indices

	mean	**min**	**max**	**std**	**SE**
ascAvFreq	13.985	13.689	14.356	0.095	0.003
ascLemmaAvFreq	10.521	8.172	11.986	0.473	0.013

Table 13. Correlations between frequency indices and L2 oral proficiency scores

	Score	**ascAvFreq**
ascAvFreq	-0.285	1.000
ascLemmaAvFreq	-0.105	0.204

4.2.1.4 SOA

For brevity, only indices with meaningful correlations ($|r| > 0.1$) and the strongest correlations within each of the four SOA sub-indices (i.e., AvMI, Tscore, DPLemmaCue, DPStructureCue) are presented (and included for the subsequent analyses in RQ 2 and RQ 3). The description statistics for SOA indices are provided in Table 14, and correlation results are detailed in Table 15. The result showed that five ASC types (intransitive resultative, passive, caused-motion, transitive resultative, transitive simple and intransitive simple) showed moderate

to small positive correlations. Conversely, intransitive motion constructions showed a moderate negative correlation.

Table 14. Descriptive statistics: SOA indices

	mean	min	max	std	SE
ascAv_DPLemmaCue	0.491	0.246	0.770	0.055	0.002
CAUS.MOT_AvMI	1.362	-6.888	4.730	1.660	0.046
DITRAN_DPLemmaCue	0.313	-0.031	0.863	0.326	0.009
INTRAN.MOT_DPStructureCue	0.167	-0.013	0.217	0.048	0.001
INTRAN.RES_Tscore	19.054	-20.925	120.354	39.425	1.102
INTRAN.S_DPStructureCue	0.009	0.062	0.000	-0.014	0.009
PASSIVE_AvMI	0.879	-7.509	5.040	1.633	0.046
TRAN.RES_AvMI	1.138	-4.483	4.745	1.790	0.050
TRAN.S_DPStructureCue	0.018	0.000	0.039	0.005	0.000

Table 15. Correlations between SOA indices and L2 oral proficiency score and SOA indices

	Score	asc	CAUS. MOT	DITRAN	INTRAN. MOT	INTRAN. RES	INTRA.S	PASSIVE	TRAN. RES
ascAvDPL	-0.134	1.000							
CAUS.MOT _AvMI	0.236	-0.021	1.000						
DITRAN _DPL	0.247	0.007	0.110	1.000					
INTRAN.MOT _DPS	-0.296	0.109	-0.112	-0.068	1.000				
INTRAN.RES _Tscore	0.434	-0.074	0.102	0.074	-0.125	1.000			
INTRAN.S _DPS	0.136	-0.010	0.002	-0.014	-0.036	0.037	1.000		
PASSIVE _AvMI	0.273	-0.016	0.076	0.125	-0.080	0.135	0.000	1.000	
TRAN.RES _AvMI	0.230	-0.023	0.066	0.068	-0.080	0.095	0.082	0.095	1.000
TRAN.S _DPS	0.189	0.291	0.098	0.032	-0.052	0.064	0.104	0.081	0.100

Notes. DPL: DPLemmaCue; DPS: DPStructureCue; The header row omits the specific sub-index type for space reasons. Each column label (e.g., CAUS.MOT, DITRAN) aligns with the corresponding sub-index in the row label (e.g., CAUS.MOT_AvMI, DITRAN_DPL).

4.2.2 RQ 2: Extent to which ASC-based indices predict L2 oral proficiency scores

For RQ 2, after filtering and multicollinearity checks, 18 indices entered the subset regression. Residuals appeared normally distributed (Q-Q plot). AIC-guided selection yielded a

final model with 14 predictors (Table 16). Figure 8 plots predicted against actual proficiency scores.

Table 16. Linear model predicting L2 oral proficiency scores using the selected ASC-based indices

	Relative importance (%)	**Estimates**	**SE**	***t***	***p***
(Intercept)		-3.179	1.389	-2.288	.022
ascAvDPLemmaCue	2.26	-4.017	0.905	-4.441	<.001
ascLemmaAvFreq	1.36	0.487	0.126	3.854	<.001
ascLemmaMATTR11NoBe	7.92	3.820	0.674	5.670	<.001
CAUS.MOT_AvMI	4.98	0.085	0.024	3.555	<.001
CAUS.MOT_Prop	3.62	7.021	3.057	2.297	.022
DITRAN_DPLemmaCue	8.60	0.790	0.103	7.638	<.001
INTRAN.MOT_DPStructureCue	9.28	-4.800	0.731	-6.567	<.001
INTRAN.RES_Tscore	27.83	0.012	0.001	13.856	<.001
INTRAN.S_DPStructureCue	2.49	10.886	4.058	2.683	.007
INTRAN.S_Prop	3.85	3.167	0.637	4.974	<.001
PASSIVE_AvMI	7.47	0.102	0.022	4.658	<.001
PASSIVE_Prop	9.50	18.560	2.679	6.928	<.001
TRAN.RES_AvMI	5.66	0.093	0.019	4.900	<.001
TRAN.S_DPStructureCue	5.20	31.123	7.230	4.305	<.001

The model's residual standard error is 1.183, with an R^2 value of 0.441 (adjusted R^2 = 0.435), indicating that approximately 44% of the variability in L2 oral proficiency scores is explained by the model. The F-statistic is 71.27 on 14 and 1266 degrees of freedom, with a *p*-value of < .001, indicating that the model is statistically significant. The correlation between the predicted scores and the actual scores, calculated as the square root of the R^2 value, is $r = \sqrt{0.441} \approx 0.664$ (strong positive correlation; Cohen, 1988).

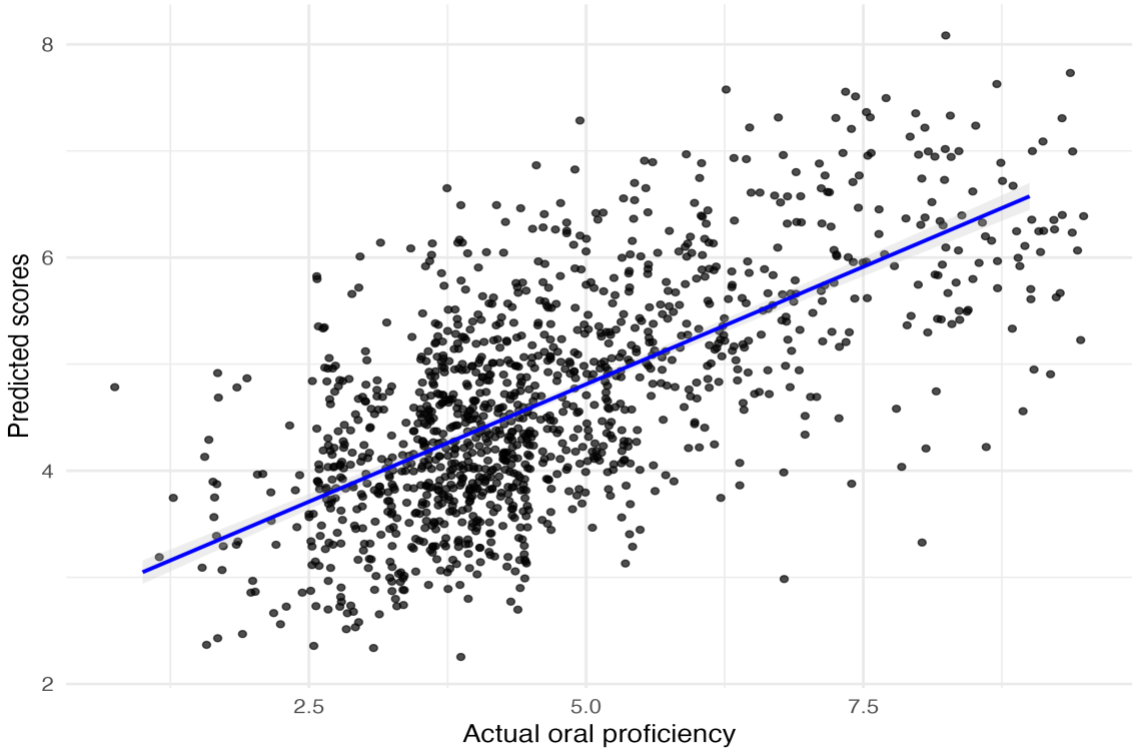

Figure 8. Actual L2 oral proficiency scores vs. scores predicted by the best model (ASC-based indices).
Notes. The blue line represents a linear fit that illustrates the overall trend between actual and predicted scores, while the grey area indicates the 95% confidence interval around the fit.

4.2.3 RQ 3: Extent to which ASC-based and other indices predict L2 oral proficiency scores

To address RQ 3, a subset of the ASC-based indices and other established lexicogrammatical indices was included in a linear regression model following statistical filtering. Prior to conducting the regression analysis, the residuals were evaluated (Q-Q plot), confirming their normality. The best model was identified using AIC, resulting in a final model comprising 17 selected indices as predictors (Table 17). Figure 9 visualizes the relationship between predicted and actual oral proficiency scores in this optimal model, illustrating a stronger alignment. Figure 10 displays each predictor's relative importance by constructs.

Table 17. Linear model predicting L2 oral proficiency scores using the selected ASC-based and other lexicogrammatical indices

	Relative importance (%)	Estimates	SE	*t*	*p*
(Intercept)		- 12.910	2.157	-5.988	<.001
ascLemmaMATTR11NoBe	3.22	2.033	0.507	4.011	<.001
b_concreteness	13.50	-1.178	0.253	-4.655	<.001
CAUS.MOT_AvMI	1.89	0.042	0.016	2.575	.010
cw_lemma_freq_log	6.46	-0.300	0.174	-1.729	.084
DITRAN_DPLemmaCue	2.47	0.360	0.083	4.317	<.001
INTRAN.MOT_DPStructureCue	3.93	-2.904	0.579	-5.016	<.001
INTRAN.RES_Tscore	9.53	0.007	0.001	9.475	<.001
INTRAN.S_Prop	1.60	1.824	0.466	3.913	<.001
lemma_bg_MI2	14.37	1.053	0.093	11.376	<.001
mattr11	9.54	8.301	1.503	5.507	<.001
noun_freq_log	7.19	0.798	0.140	5.683	<.001
PASSIVE_AvMI	2.46	0.032	0.017	1.860	.063
PASSIVE_Prop	4.46	13.190	2.105	6.264	<.001
TRAN.RES_AvMI	1.84	0.043	0.015	2.794	.005
usf	3.68	-0.024	0.006	-4.363	<.001
v_dobj_deltap_strgst	5.59	-2.282	0.390	-5.854	<.001
v_nsubj_deltap_govcue	8.30	6.537	1.074	6.088	<.001

The model exhibits a residual standard error of 0.938 and an R^2 value of 0.650 (adjusted $R^2 = 0.645$), explaining approximately 65% of the variability in L2 oral proficiency scores. The F-statistic is 137.7 on 17 and 1263 degrees of freedom, with a *p*-value of < .001, indicating statistical significance. The correlation between the predicted scores and the actual scores, calculated as the square root of the updated R^2 value, is $r = \sqrt{0.650} \approx 0.806$, which indicates a strong positive correlation.

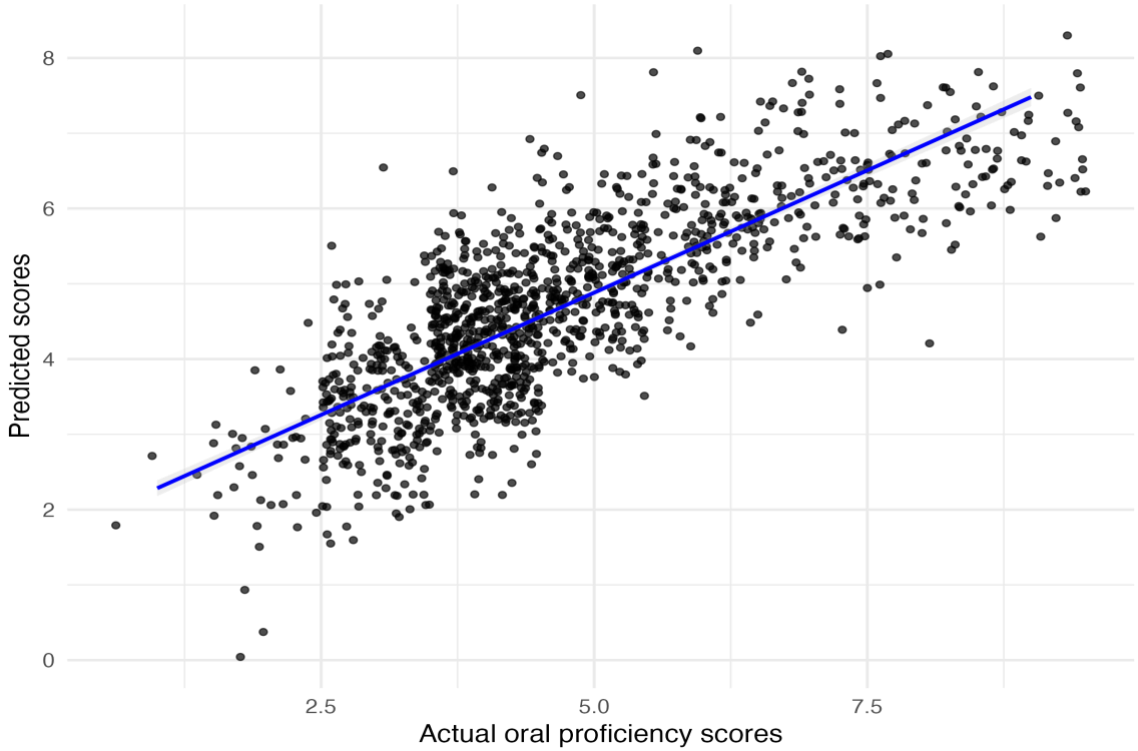

Figure 9. Actual L2 oral proficiency scores vs. scores predicted by the best model (ASC-based and other lexicogrammatical indices)

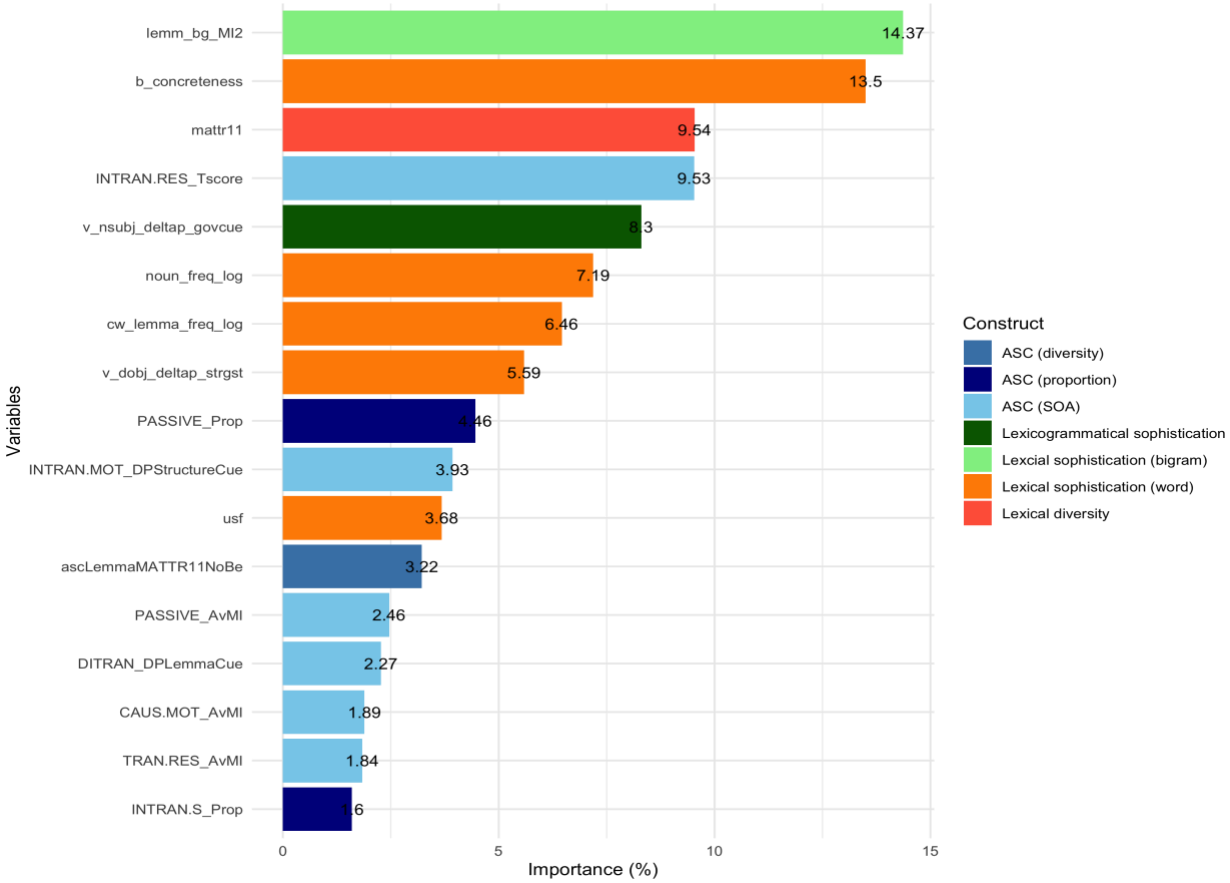

Figure 10. Summary of the relative importance of each predictor in the best model (ASC-based and other lexicogrammatical indices)

4.3 Discussion

4.3.1 RQ 1: Relationship between ASC-based indices and L2 oral proficiency scores

The overall results indicated that indices from each category exhibited small to moderate correlations with the proficiency scores. Detailed discussions for each type of index are provided below.

4.3.1.1 Diversity

The results indicated that more proficient speakers tended to use a wider variety of ASCs and ASC-verb lemma combinations, consistent with Kim and Ro's (2023) findings on advanced L2 speakers in a one-minute opinion speaking task. While Kim and Ro measured diversity using

log-transformed type frequency, a metric potentially influenced by text length (Kyle et al., 2024), the present study corroborates their findings using MATTR, which is designed to mitigate text-length effects. Notably, the same pattern was observed despite the difference in measurement approach. Furthermore, ASC-verb lemma combinations excluding the verb "be" (ascLemmaMATTR11NoBe) showed the strongest correlation with proficiency scores. This suggests that the verb "be," predominantly used in attributive constructions and often considered a basic, frequent verb in language use (Housen, 2008), may not provide significant information for modeling L2 proficiency, a notion further supported by Biber's (1988) characterization of such structures as "non-complex constructions" (p. 228) due to their lower informational content and typical presence in spoken discourse.

4.3.1.2 Proportion

The proportion indices support the finding that more proficient speakers tended to use a wider variety of constructions. Among these, intransitive resultative constructions (INTRAN.RES_Prop; e.g., "the wind **got** so strong") showed the strongest positive correlation (r = 0.307), followed by passive (PASSIVE_Prop; e.g., "I was **born** in January") and caused-motion constructions (CAUS.MOT_Prop; e.g., "I **dropped** my bag on the rail"), which also showed positive correlations (r = 0.291 and r = 0.216, respectively). These results echo Hwang and Kim's (2023) observation that higher L2 writing proficiency is associated with a greater use of complex constructions, even though the language-use domain is different from this study. Notably, whereas Hwang and Kim identified the *there*-expletive construction as the strongest predictor of L2 writing proficiency, the use of intransitive resultative constructions is highlighted in this study. This discrepancy may stem from variations in genre, task type, or learner characteristics; however, the spoken domain might also be a critical factor, as it tends to favor

succinct constructions like intransitive resultatives that support dynamic, interactive communication (Choi & Sung, 2020).

Meanwhile, the proportion of transitive simple constructions (TRAN.S_Prop; e.g., "They **explained** what happened there") exhibited the strongest negative correlation with proficiency (r = -0.215), suggesting that reliance on more frequent, thus predictable, forms may indicate limited linguistic range. Other common structures, such as attributive (ATTR_Prop; e.g., "this **is** a kind of party") and intransitive simple constructions (INTRAN.S_Prop; e.g., "what **happened** there"), showed negligible or weak correlations with proficiency (r = -0.079 and r = 0.174, respectively).

To provide an overview of the proportional indices, the L2 speakers were categorized into groups, and each group's distribution of proportion indices across proficiency levels was plotted using varying y-axis scales (Figure 11) and a consistent scale (Figure 12) to compare the ASC types. The figures show that while more proficient speakers tended to use a wider variety of constructions and leaned towards more "complex" ones, most of their speech still relied on attributive, transitive simple, and intransitive simple constructions. This suggests that the "simple" constructions remain central for both advanced and less proficient speakers, with a subtle increase in producing the "complex" constructions among the more proficient.

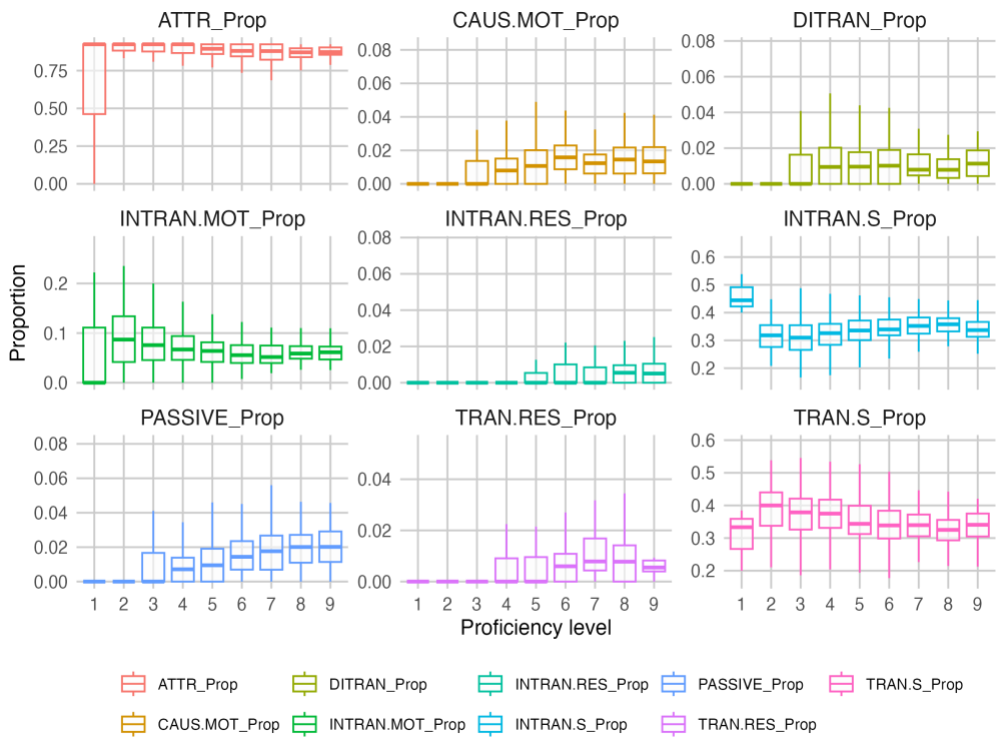

Figure 11. Distribution of proportion indices across oral proficiency levels (x-axis) with individual y-axis scales for each index

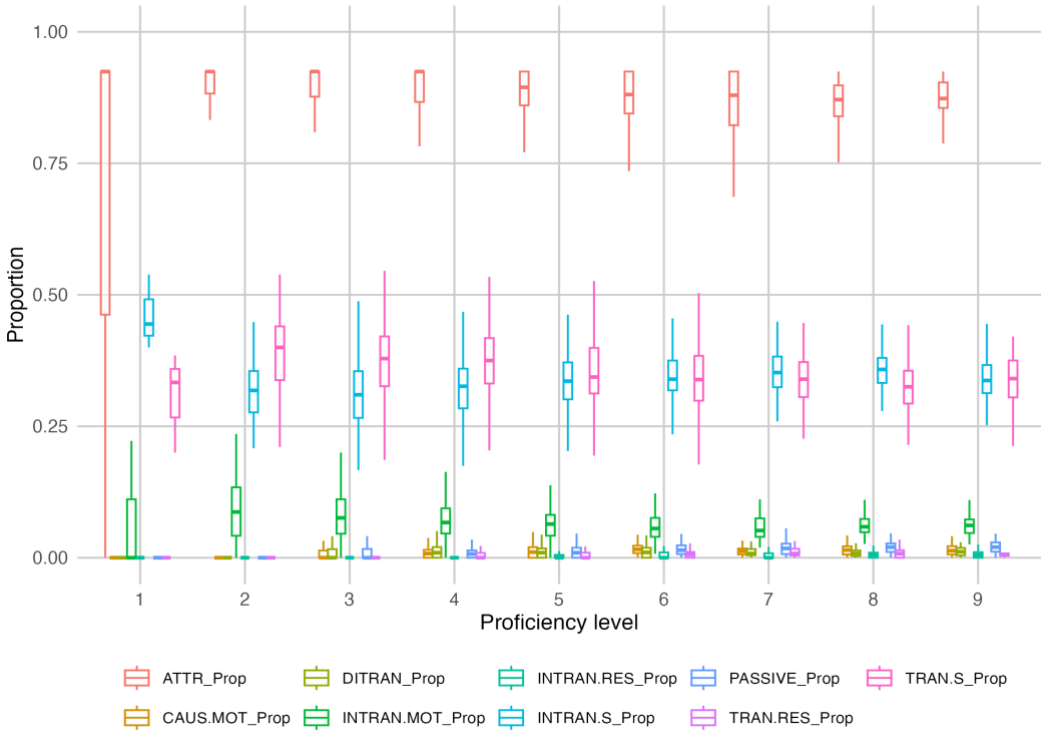

Figure 12. Unified scale comparison of proportion indices across oral proficiency levels

4.3.1.3 Frequency

The results indicated that more proficient speakers tended to use less frequent ASCs (ascAvFreq, r = -0.285) and, to a lesser extent, less frequent ASC-verb lemma combinations (ascLemmaAvFreq; r = -0.105). In particular, they favored less common ASCs, such as intransitive, ditransitive, transitive resultative, and caused-motion constructions (Table 18), which aligns with the findings from the diversity and proportion indices.

Table 18. Frequency of ASC and ASC-verb combinations from the SUBTLEX-US corpus

ASC type	Frequency	Frequent verbs (Count of verb appearances)
TRAN.S	2,280,329	*have* (143,988), *know* (135,579), *do* (133,477)
ATTR	1,458,635	*be* (1,348,583), *look* (24,136), *feel* (17,217)
PASSIVE	153,042	*do* (4,394), *call* (3,910), *kill* (3,098)
INTRAN.S	879,675	*go* (77,878), *know* (60,706), *come* (59,374)
INTRAN.MOT	313,242	*get* (18,178), *go* (6,582), *become* (6,149)
CAUS.MOT	162,262	*get* (21,410), *put* (20,971), *take* (17,691)
TRAN.RES	194,361	*let* (51,151), *make* (26,066), *want* (21,059)
DITRAN	174,268	*tell* (71,291), *give* (35413), *ask* (11768)
INTRAN.RES	49,437	*become* (198,068), *get* (71,808), *go* (35,631)

Usage-based approaches (Ellis, 2002a, b; Tomasello, 2005) suggest that *less* proficient speakers rely on verbs that inherently encode the meaning of the ASC (e.g., "become" in intransitive resultatives or "give" in ditransitive) or on frequently used light verbs such as "have," "do," and "get." While previous L2 writing studies (Kyle & Crossley, 2017; Kyle et al., 2021) found that advanced writers produce less frequent construction-verb lemma combinations, this effect appears smaller in spoken language, where advanced speakers still use many light verbs (Gilquin, 2019; Kim & Rah, 2019). For example, an advanced speaker (Level 9) said, "When I first **went**$_{\text{INTRAN.MOT}}$ to America, I **had**$_{\text{TRAN.S}}$ a problem understanding$_{\text{TRAN.S}}$ the different sizes," while a beginner speaker (Level 2) said, "We **went**$_{\text{INTRAN.MOT}}$ to movie theater. We buy$_{\text{TRAN.S}}$ popcorn and coke at movie theater," when asked by the interviewer in a similar

context. Additionally, the key difference sometimes lay in the use of more embedded ASCs (i.e., dependent clauses) with light verbs as the main verb, including non-finite clauses, which may not significantly contribute to the use of less frequent ASC-verb lemma combinations. For example, when using the same light verb "get," an advanced speaker (Level 8) produced sentences like, "I **got**$_{\text{INTRAN.MOT}}$ on the train from that station [which nobody's working$_{\text{INTRAN.S}}$ there]," whereas a beginner (Level 2) produced simpler constructions such as, "which train I **get**$_{\text{INTRAN.MOT}}$," in a similar context.

4.3.1.4 SOA

The results indicated that more proficient speakers tended to use verbs that were more strongly associated with specific ASC types, with small to medium positive correlations. This finding aligns with previous studies on VACs in L2 written production (Kyle & Crossley, 2017; Liu & Lu, 2024) and dependency bigrams in L2 spoken production (Kyle & Eguchi, 2023; Paquot, 2018; Rubin et al., 2021). This trend was evident across most ASC types, except for the intransitive motion, where the association strength decreased with proficiency.

To further analyze the SOA indices, two DeltaP indices (i.e., DeltaPStructureCue, DeltaPLemmaCue) were plotted for each proficiency level (Figures 13 and 14). Overall, more proficient speakers tended to use verbs that were more strongly associated with the listed ASC types, while the opposite pattern was consistently found for the intransitive motion construction. One possible explanation is that the verb "go," which is typically strongly associated with intransitive motion, is increasingly replaced by less conventional verbs with metaphorical uses (categorized into INTRAN.MOT; e.g., "talk to", "explain to") particularly among more proficient speakers. These metaphorical uses, in which even movement verbs sometimes express

abstract concepts (e.g., “she went into great detail about the topic”, “they walked through the reasoning step by step”) may reflect higher language proficiency (Boas, 2010; Talmy, 1985).

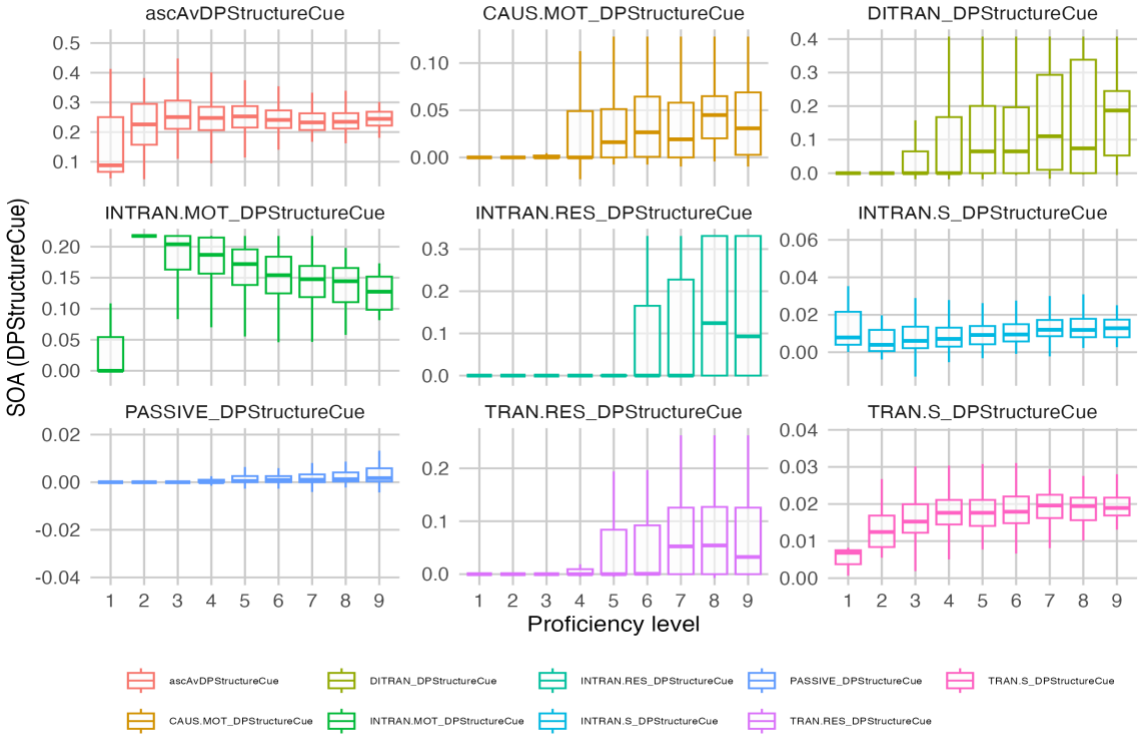

Figure 13. Distribution of SOA indices (*DeltaPStructureCue*) across oral proficiency levels with individual y-axis scales for each index

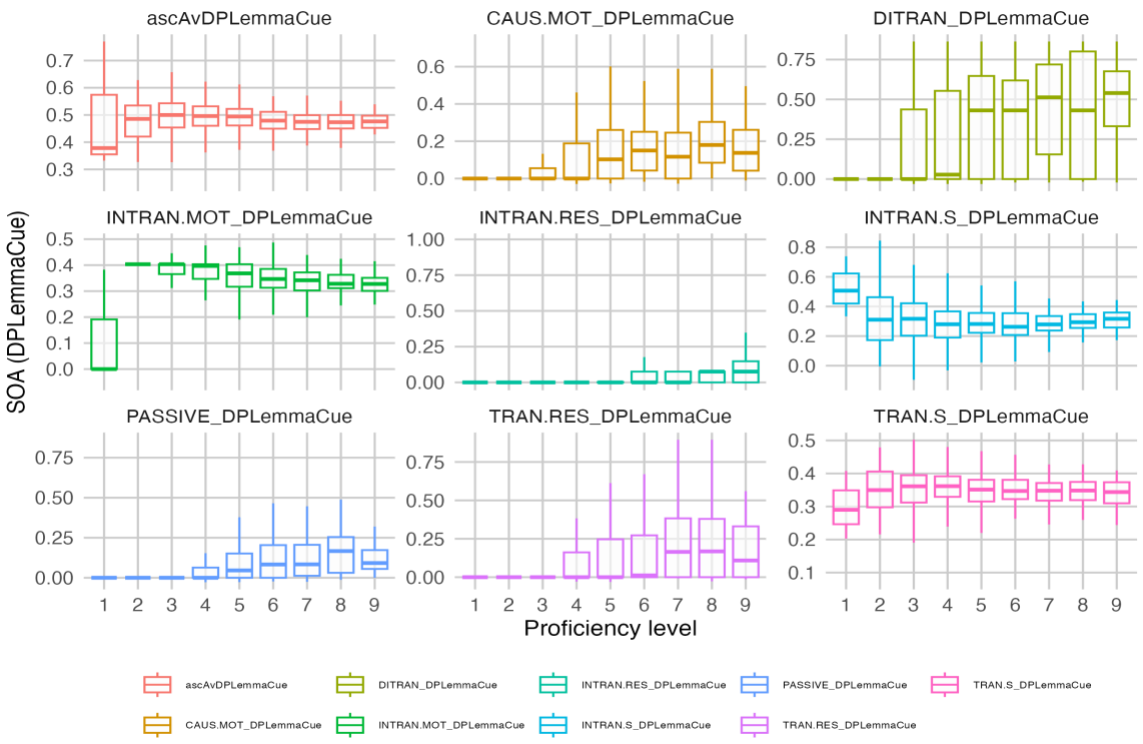

Figure 14. Distribution of SOA indices (*DeltaPLemmaCue*) across oral proficiency levels (x-axis) with individual y-axis scales for each index

4.3.2 RQ 2: Extent to which ASC-based indices predict L2 oral proficiency scores

The best multivariate linear regression model used 14 selected ASC-based indices, which collectively accounted for 44% of the variability in the scores. Further analysis showed that the SOA category, with a cumulative importance of 73.77%, played an important role in capturing lexicogrammatical aspects of language use in the L2 oral proficiency test. Proportion indices contributed 13.12% to the model, while diversity indices added 7.92%. As discussed in RQ 1, these categories emphasize the diverse and proportional use of linguistic constructions in language production. However, even advanced L2 speakers primarily relied on basic constructions (e.g., attributive, simple transitive, and simple intransitive), suggesting that the modest increase in complex constructions (e.g., caused-motion, intransitive resultative, transitive resultative, ditransitive, and passive) did not substantially account for variance in proficiency

scores. The frequency category contributed the least, at just 1.36%, indicating that while the raw frequency of ASC or ASC–verb lemma usage has some relevance, it is far less predictive than the diversity and functional integration of constructions and verbs.

Overall, this finding supports the validity of measuring the use of verb and ASCs, as highlighted in previous research (Kyle, 2016; Kyle & Crossley, 2017; Kyle et al., 2021; Hwang & Kim, 2023; Kim et al., 2018; Kim & Hwang, 2022; Kim & Ro, 2023; Li & Yu, 2024). Consistent with the usage-based approach to L2 acquisition, which emphasizes that language learning emerges from usage patterns and frequency-based exposure, the results support that L2 speakers' increasing ability to produce a diverse range of verbs and ASCs reflects their developing grammatical proficiency (Kim & Ro, 2023). In addition, this study demonstrates that SOA measures provide significant explanatory power and further insights into ASC usage in L2 spoken production (Kyle, 2016; Kyle & Crossley, 2017).

4.3.3 RQ 3: Extent to which ASC-based and other indices predict L2 oral proficiency scores

Using 17 selected indices, 65% of the variability in the scores was explained, an increase of 21 percentage points compared to the model built from RQ 2. Follow-up analysis demonstrated that the ASC-based category (31.38% importance), lexical sophistication (30.38%), bigram-related SOA indices (28.25%), and lexical diversity (9.54%), all contributed meaningfully. This finding also extends previous studies (Eguchi & Kyle, 2020; Kyle & Eguchi, 2023), which each explained approximately 59% of the variance in the L2 oral proficiency scores using lexicogrammatical indices, by capturing an additional 6 percentage points of variance. Despite the substantial proportion of explained variance, 35% remains unaccounted for, echoing the limitations noted by Kyle and Eguchi (2023), who recommended that future research

incorporate additional variables (e.g., fluency, accuracy) to further enhance model explanatory power.

In summary, the results demonstrate that the ASC-based indices not only reveal how learners employ ASCs but also provide additional explanatory power beyond established lexicogrammatical measures. By capturing clausal-level complexity and sophistication (centered on ASCs and their associated verbs), these indices complement word- and multiword-level indices, thereby enriching multivariate analyses of L2 oral proficiency. These findings also align with a growing body of research highlighting the importance of assessing productive proficiency as a multifaceted construct (Bulté & Roothooft, 2020; Eguchi & Kyle, 2020; Kim et al., 2018; Kyle & Eguchi, 2023; Saito, 2020; Yoon et al., 2019).

4.4 Summary

This study aimed to address a gap in the literature by exploring the lexicogrammatical complexity and sophistication of L2 spoken proficiency using ASC-based indices, a domain that has received less attention compared to L2 written production. Data were drawn from a corpus of 1,281 oral proficiency interviews with Japanese learners of English, and participants were assessed using a standardized speaking test adapted from the ACTFL Oral Proficiency Interview (Section 4.1.1). The study investigated the relationships among ASC-based indices, oral proficiency scores, and broader lexicogrammatical indices (Section 4.1.2) through correlation and regression analyses (Section 4.1.3). The results indicated that ASC-based indices alone accounted for 44% of the variance in oral proficiency scores, providing fine-grained insights into lexicogrammatical complexity and sophistication (Sections 4.2.1 and 4.2.2). When combined with additional lexicogrammatical measures, the model's explanatory power increased to 65% (Section 4.2.3), capturing an additional 6 percentage points compared to the model using only

the previously established lexicogrammatical indices (59% as described in Kyle & Eguchi, 2023). These findings support the validity of ASC-based indices in capturing dimensions of lexicogrammatical complexity and sophistication not accounted for by conventional measures and underscore the value of integrating diverse linguistic indices for a more comprehensive assessment of oral proficiency (Section 4.3).

5 Measuring lexicogrammatical complexity and sophistication in L2 writing assessment

Over the past few decades, research on L2 writing proficiency has shifted from traditional length-based metrics to more nuanced constructs of syntactic complexity. As discussed in Section 2.3, earlier studies questioned the empirical validity of the length-based measures (e.g., mean length of clauses) in assessing writing proficiency. In response, some researchers have adopted fine-grained constructs of syntactic complexity, with noun phrase complexity emerging as a particularly effective predictor of L2 writing proficiency (Biber et al., 2011, 2014; Kyle, 2016). At the same time, empirical studies have consistently highlighted the predictive power of length-based measures and the frequency of certain syntactic structures (e.g., coordinate phrases, complex nominals) per clause in evaluating L2 writing proficiency (Casal & Lee, 2019; Lu, 2010, 2011; Kim, 2014; Li, 2015; Yang et al., 2015).

However, this emphasis on syntactic complexity does not mean that the constructs of lexicogrammatical complexity and sophistication have been overlooked in L2 writing research. A substantial body of research has examined lexical diversity (Zenker & Kyle, 2021) and both lexical and lexicogrammatical sophistication (Eguchi & Kyle, 2023; Kyle & Eguchi, 2021). At the clausal level, there has also been a growing interest in constructional complexity and sophistication, with several studies comparing construction-based indices to writing proficiency levels (Kyle & Crossley, 2017; Kyle et al., 2021; Hwang & Kim, 2023; Kim et al., 2023). Nevertheless, these constructs have yet to be rigorously evaluated within the framework of focused ASCs, particularly regarding their interaction with verbs, nor have they been systematically tested within a multivariate framework.

Alongside addressing methodological challenges associated with extracting ASCs, similar to those identified in the study on L2 speaking assessment (Chapter 4), this chapter aims to fill outlined research gaps by exploring the following RQs:

RQ 1. What is the relationship between L2 writing proficiency scores and ASC-based indices?

RQ 2. To what extent are ASC-based indices predictive of L2 writing proficiency scores?

RQ 3. To what extent are ASC-based indices predictive of L2 oral proficiency scores when combined with other lexicogrammatical and syntactic complexity indices?

5.1 Method

5.1.1 Datasets

5.1.1.1 Learner corpus

The English Language Learning Insight, Proficiency, and Skills Evaluation (ELLIPSE) corpus (Crossley et al., 2023) was used in this study. It consists of 6,482 essays written by English language learners in an ESL setting during statewide standardized annual testing across the United States. These essays span 29 distinct independent writing prompts, each designed to require no specialized background knowledge or source texts. Each essay is accompanied by demographic information detailing the economic status, gender, grade level (grades 8-12), and race or ethnicity.

The original essays comprising the ELLIPSE corpus were selected from a larger pool of approximately 600,000 essays gathered from statewide and national standardized tests assessing students' writing skills. Essays included in the ELLIPSE corpus met three main criteria: they were explicitly labeled as written by ESL students, had comprehensive demographic and

individual difference data, and contained at least 75% correctly spelled English words. Initially, this subset included 8,890 essays across 44 distinct prompts.

To evaluate language proficiency, the project developed a specialized rubric based on a review of 18 academic articles and 38 existing proficiency rubrics. The rubric underwent iterative refinement, guided by feedback from a teacher advisory board of ten ESL educators and a research advisory board comprising experts in second language acquisition, ESL education, and composition. The final rubric includes a holistic proficiency score and six analytical scores evaluating cohesion, syntax, vocabulary, phraseology, grammar, and orthographic/punctuation conventions. Each component is rated on a 5-point Likert scale, from 1 (indicating limited English proficiency) to 5 (representing native-like proficiency) (Appendix C).

All essays in the initial corpus were independently scored by at least two trained raters. A total of 26 raters were recruited from a research university in the southeastern United States. The majority were female (n = 21) with diverse academic backgrounds (primarily in applied linguistics or English) ranging from advanced undergraduates to doctoral students, all with prior ESL teaching experience. A Many-Facet Rasch Measurement analysis was then employed to examine the reliability of the essays, raters, and scoring rubric items. Reliability metrics, analogous to Cronbach's alpha, were generated to assess consistency across texts, raters, and rubric scales. Essays, raters, or scales exhibiting extreme variability or poor consistency (with infit values outside the acceptable range of 0.6 to 1.4) were excluded. This screening process resulted in a final corpus of 6,482 essays, demonstrating robust reliability for subsequent analyses.

For this study, an averaged of *syntax*, *vocabulary*, *phraseology*, and *grammar* scores was used to measure L2 writing proficiency, as these constructs closely reflect the target domain of lexicogrammatical complexity and sophistication.

5.1.1.2 Reference corpus

The English-Corpus of Web (EnCOW; Schäfer, 2015; Schäfer & Bildhauer, 2012) is a large-scale, web-harvested corpus of English, used as a benchmark for advanced writing. In this study, a subset of EnCOW comprising 360,783,433-word tokens, 15,439,673 sentences, and 39,838,785 tagged ASCs was used. Figure 15 illustrates the normalized distribution of ASC types in the ELLIPSE learner corpus versus the EnCOW reference corpus. Figure 15 presents the normalized distribution of various ASCs across the learner and the reference corpora, illustrating how the frequency of ASC usage differs between the two datasets.

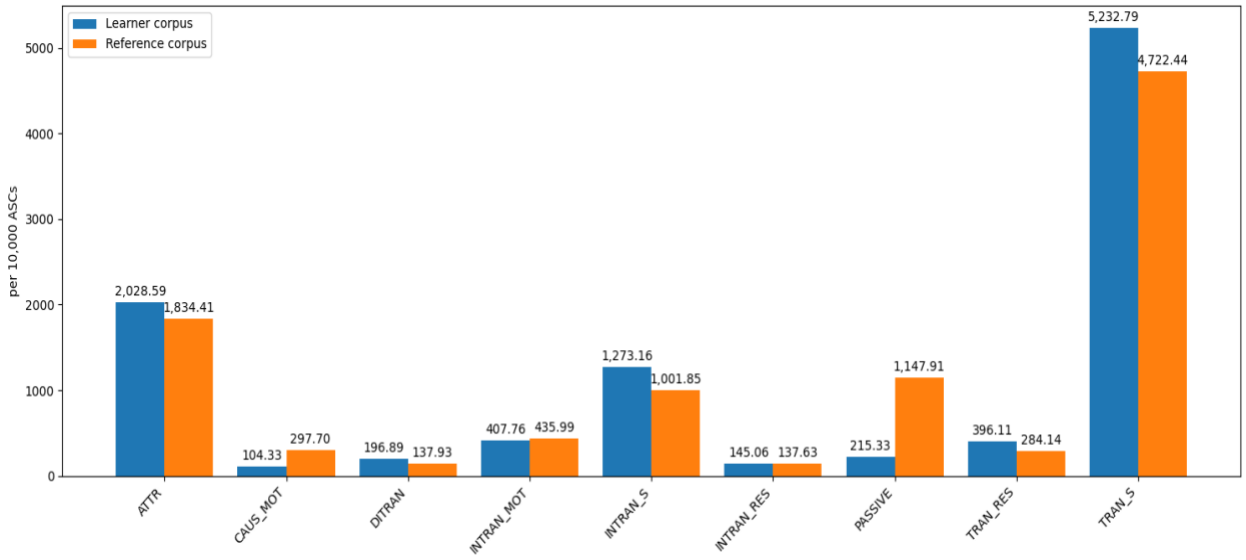

Figure 15. Distribution of ASCs in the ELLIPSE (learner) and EnCOW (reference) corpora
Notes. Frequency counts normalized by occurrences per 10,000 ASCs

5.1.2 Additional construct and indices: Syntactic complexity

L2 writing proficiency was evaluated with a multivariate framework that integrates indices of lexicogrammatical complexity and sophistication alongside measures of syntactic complexity. Following the earlier study on ASC use and L2 oral proficiency in Chapter 4, lexicogrammatical complexity and sophistication was operationalized via a range of indices (see Section 4.1.2.1-4.1.2.4). Syntactic complexity was incorporated as an additional construct in this framework because of its well-established *predictive* validity and robust empirical support in L2 writing research (Casal & Lee, 2019; Lu, 2010, 2011; Kim, 2014; Kyle, 2016; Kyle & Crossley, 2017; Li, 2015; Yang et al., 2015). This construct was operationalized through text-internal syntactic complexity indices, grouped into three categories based on the structures they quantify: length-based measures, clausal complexity indices, and phrasal complexity indices.

Length-based measures capture average clause and T-unit length, with mean length of clause (*mlc*) calculated as total words divided by the number of finite clauses, and mean length of T-unit (*mltu*) indicating the average words per T-unit—defined as an independent clause plus its dependent clauses (Hunt, 1965).[14] Clausal complexity indices quantify the frequency and distribution of embedded constructions, including dependent clauses per clause (*dc_c*), finite complement clauses per clause (*ccomp_c*), and finite relative clauses per clause (*relcl_c*), as well as the proportions of infinitive and nonfinite structures (*infinitive_prop*, *nonfinite_prop*). Phrasal complexity is assessed by examining noun phrase elaboration (cf. Biber et al., 2011, 2016): *mean_nominal_deps* divides the total number of nominal dependents by the total number of nominals to yield a global index of noun-phrase density, while fine-grained modifiers—relative

[14] Because T-units may contain multiple clauses, *mlc* and *mltu* can exhibit collinearity.

clauses (*relcl_nominal*), adjectival modifiers (*amod_nominal*), determiners (*det_nominal*), prepositional modifiers (*prep_nominal*), possessives (*poss_nominal*), and coordinating conjunctions (*cc_nominal*)—reveal which specific noun-phrase elaborations L2 writers employ most frequently. Taken together, these length-based, clausal, and phrasal indices offer a comprehensive multivariate view of L2 writers' syntactic development, elucidating not only how learners extend clauses or T-units but also how they refine internal phrase structure (cf. Kyle, 2016). Table 19 summarizes all investigated constructs, their sub-constructs, and representative indices, for measuring L2 writing proficiency.

Table 19. Summary the target constructs and related indices in measuring L2 writing proficiency

Target construct	**Sub-construct**	**Example indices**	**Related Section**
Lexicogrammatical complexity	ASC diversity	*ascMATTR11*	§3.4
	ASC proportion	*ATTR_Prop*	
Lexicogrammatical sophistication	ASC frequency	*ascAvFreq, ascLemmaFreq*	
	ASC SOA	*ATTR_AvMI*	
	dependency bigram	*n_nnmod_MI, v_advmod_T*	§4.1.2.1
Lexical sophistication	bigram usage	*lemma_bg_MI, raw_bg_MI*	§4.1.2.2-§4.1.2.3
	word frequency	*cw_lemma_freq_log*	
	concreteness	*b_concreteness*	
	contextual distinctiveness	*mcd, usf*	
Lexical diversity	word variety	*MATTR11*	§4.1.2.4
Syntactic complexity	unit length	*mlc, mltu*	§5.1.2
	clausal complexity	*dc_c, ccomp_c, relcl_c*	
	noun phrase complexity	*mean_nominal deps*	

5.1.3 Statistical analyses

To address RQ 1, correlation analyses were conducted to examine relationships between each ASC-based index and L2 writing proficiency scores. Distributions were inspected for skewness and exclusion criteria applied, after which Pearson's *r* was calculated for the remaining indices. For RQ 2 and RQ 3, the same statistical-filtering and regression-modeling procedures

used in the spoken proficiency analyses were employed: variables with negligible correlations ($|r| < .10$) were excluded; SOA indices with the strongest correlations were retained; and multicollinearity was evaluated via pairwise correlations ($|r| > .90$) and Variance Inflation Factors (VIF > 5). Candidate regression models were compared using the Akaike Information Criterion (AIC) via the *MuMIn* package's *dredge* function (Barton, 2023), retaining only those with $\Delta AIC < 4$ and selecting the model with the lowest AIC for interpretation. Full details of the statistical procedures and software are provided in Section 4.1.3.

5.2 Results

5.2.1 RQ 1: Relationship between ASC-based indices and L2 writing proficiency scores

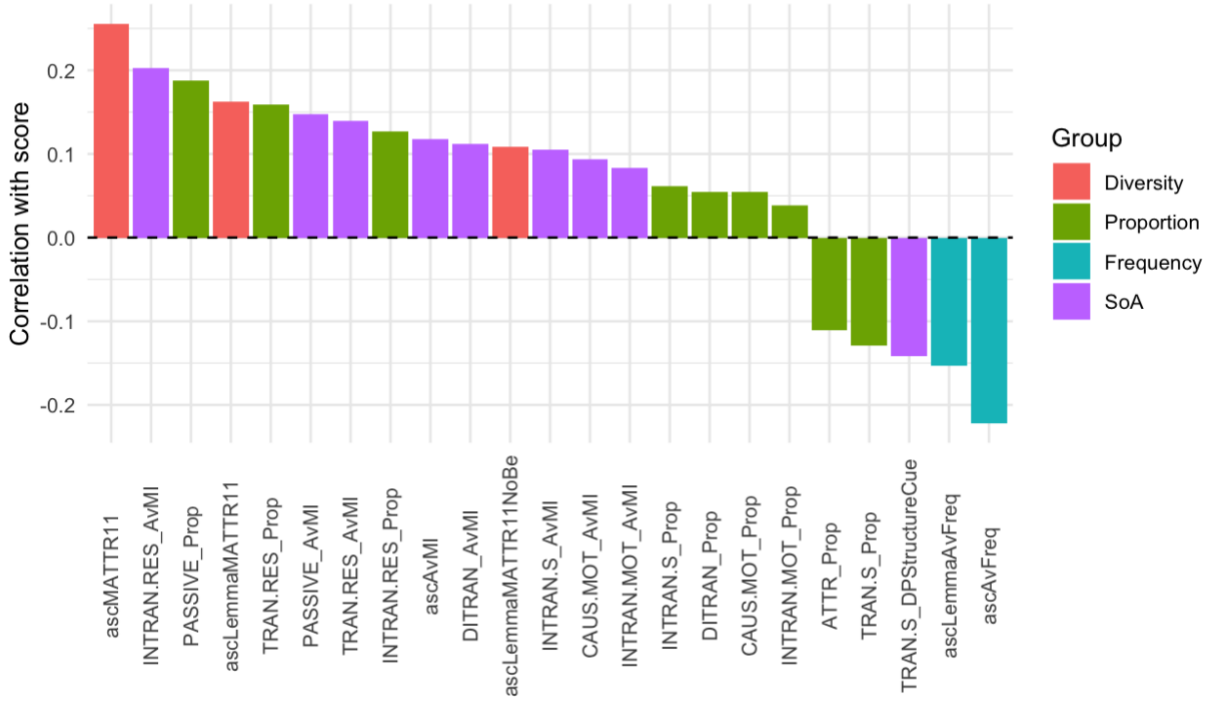

Figure 16. Correlations between L2 writing proficiency scores and selected indices of ASC use

To address RQ 1, descriptive statistics and correlations between writing proficiency scores and each index are presented. The overview results for each index group: diversity, proportion, frequency, and SOA were presented in Figure 16, with detailed results for each index type following below.

5.2.1.1 Diversity

Table 20 presents the descriptive statistics for the diversity indices. The correlation results between the indices and proficiency scores are summarized in Table 21. The findings reveal that all examined indices are significantly correlated with writing proficiency, though the strength of these correlations is generally weak.

Table 20. Descriptive statistics: Diversity indices

	mean	**min**	**max**	**std**	**SE**
ascMATTR11	0.362	0.139	1.000	0.047	0.001
ascLemmaMATTR11	0.763	0.273	1.000	0.067	0.001
ascLemmaMATTR11NoBe	0.795	0.273	1.000	0.067	0.001

Table 21. Correlations between diversity indices and L2 writing proficiency scores

	Score	**ascMATTR11**	**ascLemmaMATTR11**
ascMATTR11	0.255	1.000	
ascLemmaMATTR11	0.163	0.354	1.000
ascLemmaMATTR11NoBe	0.109	0.362	0.806

5.2.1.2 Proportion

The descriptive statistics for the proportion indices are presented in Table 22, and Table 23 shows the correlations between each index and writing proficiency scores. All indices correlated significantly with proficiency, though effect sizes were small. Attributive constructions showed a negative correlation (r = -0.111), as did transitive simple constructions

(r = -0.130). Passive constructions exhibited the largest positive correlation (r = 0.188), followed by transitive resultative (r = 0.159) and intransitive resultative constructions (r = 0.127).

Table 22. Descriptive statistics: Proportion indices

	mean	min	Max	std	SE
ATTR_Prop	0.184	0.000	0.615	0.076	0.001
CAUS.MOT_Prop	0.013	0.000	0.241	0.019	0.000
DITRAN_Prop	0.016	0.000	0.182	0.022	0.000
INTRAN.MOT_Prop	0.035	0.000	0.292	0.033	0.000
INTRAN.RES_Prop	0.012	0.000	0.150	0.017	0.000
INTRAN.S_Prop	0.243	0.000	0.557	0.069	0.001
PASSIVE_Prop	0.019	0.000	0.223	0.025	0.000
TRAN.RES_Prop	0.035	0.000	0.238	0.034	0.000
TRAN.S_Prop	0.443	0.080	0.852	0.090	0.001

Table 23. Correlations between proportion indices and L2 writing proficiency scores

	Score	ATTR	CAUS. MOT	DITRAN	INTRAN. MOT	INTRAN. RES	INTRAN.S	PASSIVE	TRAN. RES
ATTR	-0.111	1.000							
CAUS.MOT	0.055	-0.103	1.000						
DITRAN	0.055	-0.046	0.088	1.000					
INTRAN.MOT	0.039	-0.112	0.018	0.000	1.000				
INTRAN.RES	0.127	-0.095	-0.010	-0.018	-0.009	1.000			
INTRAN.S	0.062	-0.260	-0.111	-0.152	-0.042	-0.007	1.000		
PASSIVE	0.188	-0.072	0.054	-0.069	-0.034	0.033	-0.045	1.000	
TRAN.RES	0.159	-0.137	0.060	0.019	-0.114	0.070	-0.103	0.079	1.000
TRAN.S	-0.130	-0.479	-0.107	-0.094	-0.190	-0.131	-0.424	-0.202	-0.192

5.2.1.3 Frequency

The descriptive statistics for the frequency indices are provided in Table 24. Table 25 summarizes the correlations between each index and writing proficiency scores. Both ascAvFreq (r = -0.222) and ascLemmaAvFreq (r = -0.153) were significantly negatively correlated with proficiency, with effect sizes in the small range.

Table 24. Descriptive statistics: Frequency indices

	mean	min	max	std	SE
ascAvFreq	15.857	15.186	16.520	0.173	0.002
ascLemmaAvFreq	11.986	8.405	14.834	0.544	0.007

Table 25. Correlations between frequency indices and L2 writing proficiency scores

	Score	ascAvFreq
ascAvFreq	-0.222	1.000
ascLemmaAvFreq	-0.153	0.233

5.2.1.4 SOA

For conciseness, only indices with the strongest correlations within each of the four SOA sub-indices (AvMI, Tscore, DPLemmaCue, DPStructureCue) are presented and included in the subsequent analyses for RQ 2 and RQ 3. The descriptive statistics for the SOA indices are provided in Table 26, and the correlation results are detailed in Table 27. The findings indicate that intransitive resultative (r = 0.203), passive (r = 0.148), and transitive resultative (r = 0.139) constructions showed small to moderate positive correlations with proficiency. In contrast, transitive simple constructions showed a small negative correlation (r = -0.142) with proficiency.

Table 26. Descriptive statistics: SOA indices

	mean	min	max	std	SE
ascAvMI	1.365	-0.008	2.617	0.270	0.003
CAUS.MOT_AvMI	0.905	-7.274	4.807	1.574	0.020
DITRAN_AvMI	2.394	-7.450	5.776	2.458	0.031
INTRAN.MOT_AvMI	2.047	-8.883	4.301	1.717	0.021
INTRAN.RES_AvMI	1.924	-9.247	6.181	2.337	0.029
INTRAN.S_AvMI	1.145	-7.122	3.294	0.935	0.012
PASSIVE_AvMI	0.309	-8.720	3.123	1.040	0.013
TRAN.RES_AvMI	2.344	-7.046	5.026	1.608	0.020
TRAN.S_DPStructureCue	0.013	-0.004	0.057	0.006	0.000

Table 27. Correlations between SOA indices and L2 writing proficiency scores

	Score	asc	CAUS. MOT	DITRAN	INTRAN. MOT	INTRAN. RES	INTRAN. S	PASSIVE	TRAN. RES
asc_AvMI	0.118	1.000							
CAUS.MOT _AvMI	0.094	0.101	1.000						
DITRAN __AvMI	0.112	0.226	0.110	1.000					
INTRAN.MOT __AvMI	0.083	0.214	-0.112	0.033	1.000				
INTRAN.RES _ AvMI	0.203	0.207	0.102	0.043	0.053	1.000			
INTRAN.S _ AvMI	0.105	0.447	0.002	0.058	0.085	0.091	1.000		
PASSIVE _AvMI	0.148	0.094	0.076	-0.002	0.013	0.062	0.038	1.000	
TRAN.RES _AvMI	0.139	0.191	0.066	0.020	0.013	0.024	0.056	0.017	1.000
TRAN.S _DPS	-0.142	0.086	0.098	0.023	0.100	-0.002	0.085	-0.081	-0.093

5.2.2 RQ 2: Extent to which ASC-based indices predict L2 writing proficiency scores

Following the statistical filtering and addressing multicollinearity, 24 indices were retained for a subset-selection linear regression model. Residual normality was confirmed via a Q-Q plot before analysis. Model comparison using AIC identified the optimal model, which included 12 predictors (Table 28). Figure 17 plots the relationship between the predicted versus observed writing proficiency scores for this model.

Table 28. Linear model predicting L2 writing proficiency scores using the selected ASC-based indices

	Relative importance (%)	Estimates	SE	*t*	*p*
(Intercept)		3.001	0.106	28.256	<.001
ascMATTR11	16.42	0.892	0.204	4.369	<.001
ATTR.Prop	8.14	-0.902	0.106	-8.482	<.001
DITRAN_AvMI	4.20	0.013	0.003	4.578	<.001
INTRAN.RES_AvMI	15.16	0.036	0.004	9.954	<.001
INTRAN.RES_Prop	3.42	-1.082	0.502	-2.154	0.031
INTRAN.S_AvMI	6.10	0.052	0.007	7.323	<.001
PASSIVE_AvMI	9.22	0.052	0.006	8.101	<.001
PASSIVE_Prop	12.20	2.242	0.294	7.621	<.001
TRAN.RES_AvMI	5.72	0.023	0.005	5.093	<.001
TRAN.RES_Prop	5.92	0.650	0.240	2.711	0.007
TRAN.S_DPStructureCue	7.80	-8.372	1.174	-7.131	<.001
TRAN.S_Prop	5.68	-0.518	0.101	-5.141	<.001

The model's residual standard error is 0.521, and $R^2 = 0.145$ (adjusted $R^2 = 0.143$), indicating that it explains 14.5% of the variance in L2 writing proficiency scores. The overall F-statistic ($F(12, 6469) = 91.12$, $p < .001$) confirms the model's significance. The correlation between predicted and observed scores, calculated as the square root of the R^2 value, is approximately $r \approx 0.381$, which corresponds to a moderate positive effect size (Cohen, 1988).

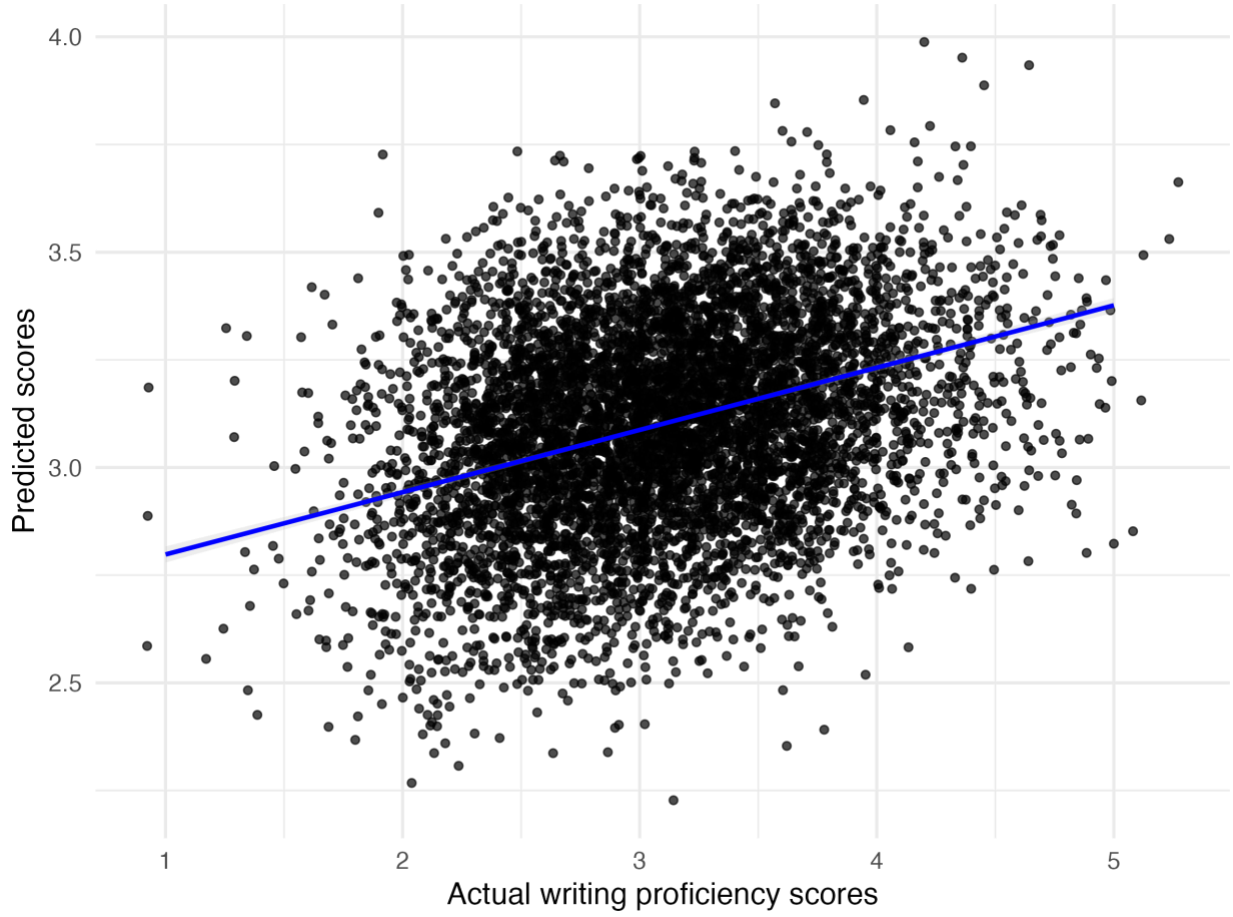

Figure 17. Actual L2 writing proficiency scores vs. scores predicted by the best model (ASC-based indices)

5.2.3 RQ 3: Extent to which ASC-based and other indices predict L2 writing proficiency scores

In addressing RQ 3, ASC-based indices were combined with the established lexicogrammatical and syntactic indices to evaluate their joint predictive power and their

incremental contribution. Table 29 summarizes the model comparisons. When used alone, ASC-based indices yielded an adjusted R^2 of 0.143; the syntactic-indices baseline achieved .077. Adding ASC-based indices to syntactic complexity increased adjusted R^2 to 0.192 (Δ = +.115). The lexicogrammatical baseline produced .363, which rose to .390 (Δ = +.027) once ASC-based indices were included. The full model, integrating ASC-based, lexicogrammatical, and syntactic indices, reached the highest explanatory power with adjusted R^2 = .407 (Δ = +.044, when compared to the lexicogrammatical-indices baseline).

Table 29. Comparison of adjusted R^2 and incremental gains by included indices

Model	**adjusted R^2**	**Δ adjusted R^2** (compared to *baseline*)
ASC-based indices	.143	
syntactic indices (*baseline 1*)	.077	
syntactic + ASC-based indices	.192	+.115
lexicogrammatical indices (*baseline 2*)	.363	
lexicogrammatical + ASC-based indices	.390	+.027
lexicogrammatical + ASC-based + syntactic indices	.407	+.044

5.2.3.1 Contribution of ASC-based indices to the syntactic complexity model

Table 30 further details the effects and relative importance of the predictors in the combined model of syntactic complexity and ASC-based indices. Among the ASC-based indices, ascMATTR11 emerged as particularly influential (relative importance = 15.15%), with a positive and statistically significant coefficient (Est. = 1.600, SE = 0.194, $p < .001$). By contrast, ccomp_c exhibited a notable negative effect (Est. = -0.015, SE = 0.001, $p < .001$) and accounted for 12.25% of the explained variance. Several other indices also contributed significantly (e.g., PASSIVE_Prop [8.78%], mltu [9.84%], and PASSIVE_AvMI [5.75%]), although none individually matched the predictive strength of ascMATTR11.

Table 30. Linear model predicting L2 writing proficiency scores using ASC-based and syntactic complexity indices

	Relative importance (%)	Estimates	SE	*t*	*p*
(Intercept)		2.800	0.107	26.262	<.001
amod_nominal	2.04	0.006	0.002	3.631	<.001
ascMATTR11	15.15	1.600	0.194	8.267	<.001
ATTR.Prop	4.13	-0.558	0.109	-5.110	<.001
ccomp_c	12.25	-0.015	0.001	-10.485	<.001
det_nominal	7.31	-0.007	0.001	-6.745	<.001
INTRAN.RES_AvMI	7.76	0.025	0.004	6.897	<.001
INTRAN.RES_Prop	2.04	-1.200	0.485	-2.477	0.013
INTRAN.S_AvMI	3.69	0.045	0.007	6.604	<.001
mltu	9.84	0.000	0.000	-5.774	<.001
nonfinite_prop	3.31	0.002	0.001	1.432	0.152
PASSIVE_AvMI	5.75	0.044	0.006	7.013	<.001
PASSIVE_Prop	8.78	2.040	0.282	7.219	<.001
TRAN.RES_AvMI	3.40	0.016	0.004	3.568	<.001
TRAN.RES_Prop	5.66	1.130	0.236	4.782	<.001
TRAN.S_DPStructureCue	5.40	-8.170	1.150	-7.139	<.001
TRAN.S_Prop	3.64	-0.243	0.100	-2.436	0.015

5.2.3.2 Contribution of ASC-based indices to lexicogrammatical model

Integrating ASC-based indices with the established lexicogrammatical indices yielded an adjusted R^2 of .390, surpassing the .363 obtained with conventional lexicogrammatical indices alone. Table 31 details the effects and relative importance of each predictor in the combined model. Figure 18 shows a strong alignment between observed and predicted proficiency scores, and the correlation of $r \approx .626$ (the square root of R^2) indicates a robust relationship between predicted and actual scores.

Table 31. Linear model predicting L2 writing proficiency scores using ASC-based and other lexicogrammatical indices

	Relative importance (%)	Estimates	SE	*t*	*p*
(Intercept)		1.598	0.856	1.867	0.062
adv_manner_freq_log	6.03	0.013	0.001	10.907	<.001
ascAvFreq	3.56	-0.309	0.047	-6.571	<.001
ascAvMI	0.94	-0.126	0.032	-3.891	<.001
ascLemmaMATTR11	1.38	-0.381	0.098	-3.873	<.001
DITRAN_AvMI	1.06	0.010	0.002	4.053	<.001
INTRAN.RES_AvMI	3.15	0.014	0.003	5.447	<.001

INTRAN.S_AvMI	1.51	0.045	0.007	6.396	<.001
mattr11	15.03	5.792	0.341	17.003	<.001
mcd	2.29	0.112	0.063	1.776	0.076
n_amod_MI	2.31	0.013	0.006	2.402	0.016
n_nnmod_MI	3.00	0.009	0.002	4.183	<.001
PASSIVE_AvMI	1.76	0.023	0.005	4.232	<.001
PASSIVE_Prop	3.46	1.627	0.235	6.936	<.001
raw_bg_MI	40.89	0.603	0.019	31.529	<.001
TRAN.RES_AvMI	1.15	0.010	0.004	2.597	0.009
TRAN.RES_Prop	1.63	0.591	0.201	2.937	0.003
usf	5.18	-0.003	0.001	-5.144	<.001
v_advmod_MI	1.69	0.018	0.008	2.200	0.028
v_dobj_MI	4.00	0.024	0.008	3.048	0.002

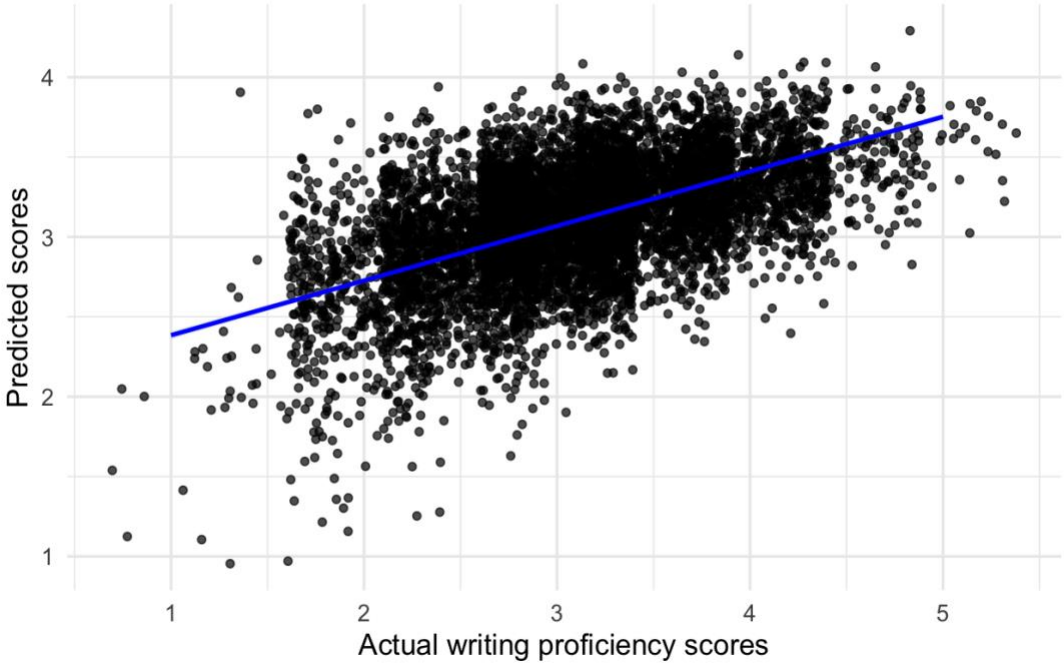

Figure 18. Actual L2 writing proficiency scores vs. Scores predicted by the best model (ASC-based and other lexicogrammatical indices)

5.2.3.3 Contribution of ASC-based, lexicogrammatical, and syntactic indices in the full model

Extending the approach from the previous two models, a final regression model was constructed to incorporate ASC-based, lexicogrammatical, and syntactic complexity indices

simultaneously. This comprehensive model yielded the highest explanatory power, achieving an adjusted R^2 of .407, an improvement of 4.4 percentage points over the baseline using only lexicogrammatical indices (adjusted R^2 = .363) and 1.7 percentage points above the model that combined ASC-based with other lexicogrammatical indices (adjusted R^2 = .390). These results underline the synergistic effect of capturing multiple facets of linguistic complexity: while lexical diversity and bigram association strength (e.g., mattr, raw_bg_MI) remain central predictors, the inclusion of ASC-based indices and syntactic complexity boosted the model's overall predictive accuracy. Table 32 details the parameter estimates and relative importance of each predictor in the full model, revealing that certain syntactic complexity indices (e.g., ccomp_c) and ASC-based measures (e.g., ascMATTR11) gain prominence when evaluated alongside established lexicogrammatical features. Figure 19 summarizes the relative importance of each predictor, grouping them by constructs and highlighting the features that contribute most significantly to writing proficiency.

Table 32. Linear model predicting L2 writing proficiency scores using ASC-based, lexicogrammatical, syntactic complexity indices

	Relative importance (%)	**Estimates**	**SE**	***t***	***p***
(Intercept)		-1.330	0.866	-1.532	0.126
adv_manner_freq_log	4.53	0.010	0.001	7.761	<.001
amod_nominal	1.10	-0.003	0.001	-3.544	<.001
ascAvFreq	2.91	-0.134	0.049	-2.709	0.007
ascLemmaMATTR11	1.20	-0.241	0.093	-2.596	0.009
ascMATTR11	4.02	0.541	0.203	2.668	0.008
ccomp_c	4.95	-0.011	0.001	-9.256	<.001
INTRAN.RES_AvMI	2.35	0.006	0.003	2.518	0.012
INTRAN.S_AvMI	1.29	0.028	0.006	4.655	<.001
mattr11	14.51	5.860	0.335	17.490	<.001
mltu	4.05	0.000	0.000	-4.285	<.001
n_nnmod_MI	2.58	0.006	0.002	2.951	0.003
PASSIVE_AvMI	1.52	0.016	0.005	3.074	0.002
PASSIVE_Prop	2.84	1.580	0.243	6.487	<.001
raw_bg_MI	38.81	0.593	0.019	31.738	<.001

TRAN.RES_Prop	2.01	1.030	0.189	5.438	<.001
usf	5.78	-0.004	0.001	-7.566	<.001
v_advmod_MI	1.65	0.017	0.008	2.142	0.032
v_dobj_MI	3.84	0.024	0.008	3.188	0.001

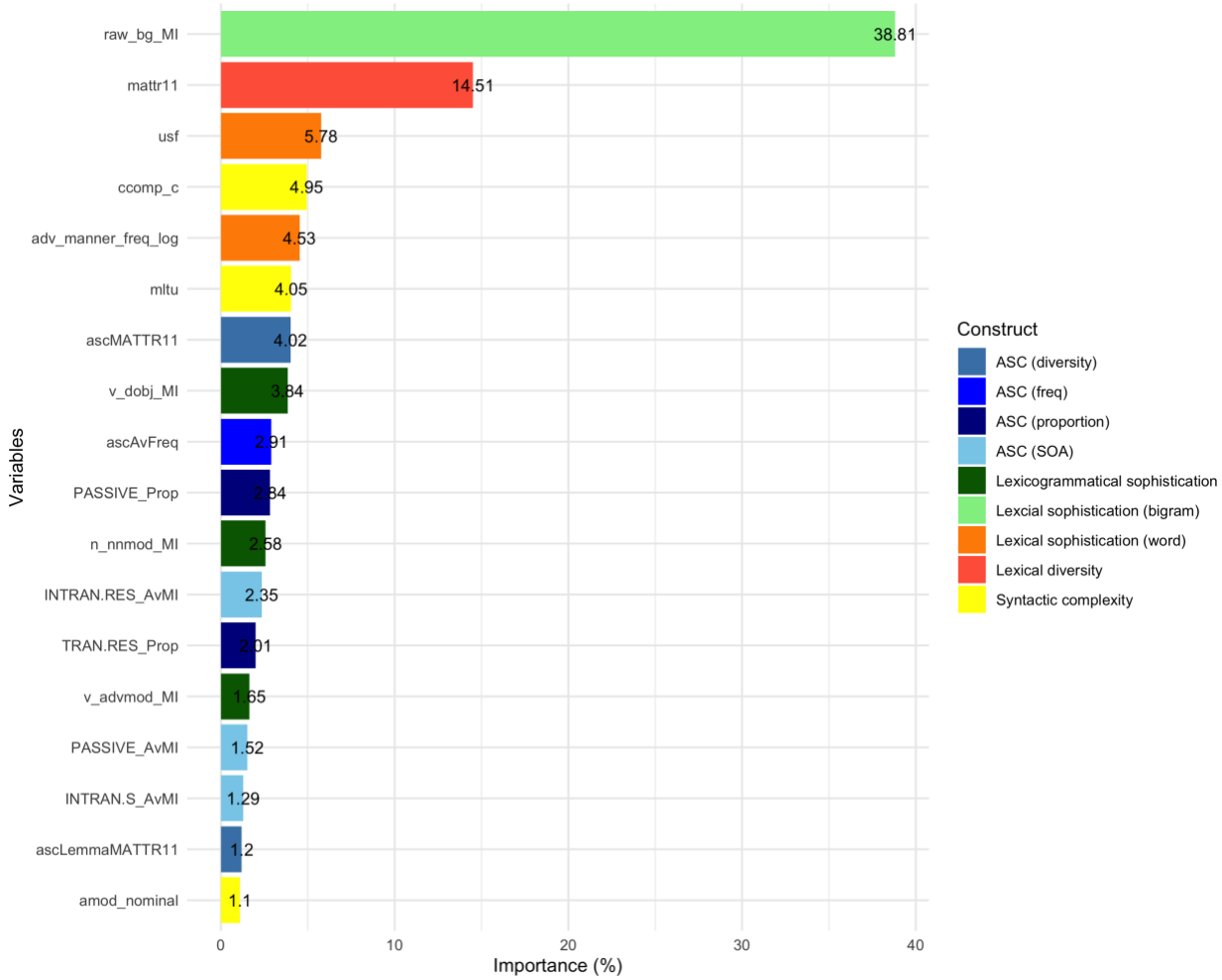

Figure 19. Summary of the relative importance of each predictor in the best model (ASC-based, lexicogrammatical, syntactic complexity indices)

5.3 Discussion

5.3.1 RQ 1: Relationship between ASC-based indices and L2 writing proficiency scores

5.3.1.1 Diversity

The results show that more proficient writers tended to use a wider range of ASCs and ASC-verb lemma combinations, aligning with previous findings that constructional diversity

rises with proficiency (Hwang & Kim 2023; Li & Yu 2024; Liu & Lu 2024). Of all indices, ascMATTR11 showed the highest correlation with proficiency, underscoring the value of varied constructional choices in effective writing. Notably, ascMATTR11 demonstrated a stronger correlation than ascLemmaMATTR11, which accounts for verb lemma variation within constructions. This suggests that in writing, the diversity of constructions themselves, regardless of the specific verbs used, is more central to signaling proficiency. This pattern was not noticeable in the spoken domain (i.e., both ascMATTR11 and ascLemmaMATTR11 correlated with proficiency scores to a similar degree). This contrast may point to modality-specific demands, whereas writing allows for more deliberate planning and encourages structural diversity, speaking relies more evenly on both syntactic patterns and lexical variation within familiar constructions to maintain fluency under time pressure.

5.3.1.2 Proportion

Of the nine ASC types investigated, five ASCs (i.e., attributive, intransitive resultative, passive, transitive resultative, and transitive simple constructions) showed significant correlations with writing proficiency scores, which partly aligns with prior research, where more proficient L2 writers tended to produce resultative and passive constructions (Hwang & Kim, 2023). In particular, the role of passives echoes Wolfe-Quintero et al.'s (1998) review, which cites Kameen's (1979) finding that stronger writers used more passives than weaker ones.

Motion constructions (i.e., intransitive motion, caused-motion), by contrast, did not show any proficiency effect; this is noteworthy given that such constructions are often considered challenging for learners from typologically distant L1 backgrounds and are viewed as metaphorical extensions of resultatives (Choi & Bowerman, 1991; Sung, 2019; Talmy, 1985). Meanwhile, attributive and transitive simple constructions correlated negatively with

proficiency, mirroring observations in the spoken task (see Section 4.3.1.2) in which more advanced L2 learners relied less on simpler constructions.

To visualize how ASCs' proportions vary across proficiency scores, scores were divided into nine equal-sized, percentile-based groups (Group 1 = lowest 10 %; Group 9 = highest 10 %).[15] The distribution of ASCs was examined through separate plots with varying y-axes for clarity (Figure 21). Quantile breakpoints guaranteed roughly equivalent sample sizes, and separate plots with individual y-axes were used for clarity (Figure 20). The visualization confirms that, while advanced writers expand their constructional repertoire and increasingly deploy complex patterns, they still draw heavily on basic grammatical templates—particularly transitive, attributive, and intransitive simple constructions. These "workhorse" structures therefore remain central even at higher proficiency levels. A useful next step would be to inspect finer-grained variation within these simple constructions (e.g., distinguishing between VERB + nominal complement and VERB + clausal complement) as such patterns may signal additional, less obvious construction types (cf. Choi & Sung, 2020; Kyle, 2016; Park & Sung, 2024). This type of analysis could help clarify whether learners gradually diversify not only beyond core grammatical templates but also within them.

[15] Group sizes (number of learners): Group 1 (740 learners), Group 2 (936 learners), Group 3 (946 learners), Group 4 (494 learners), Group 5 (976 learners), Group 6 (591 learners), Group 7 (501 learners), Group 8 (679 learners), Group 9 (619 learners).

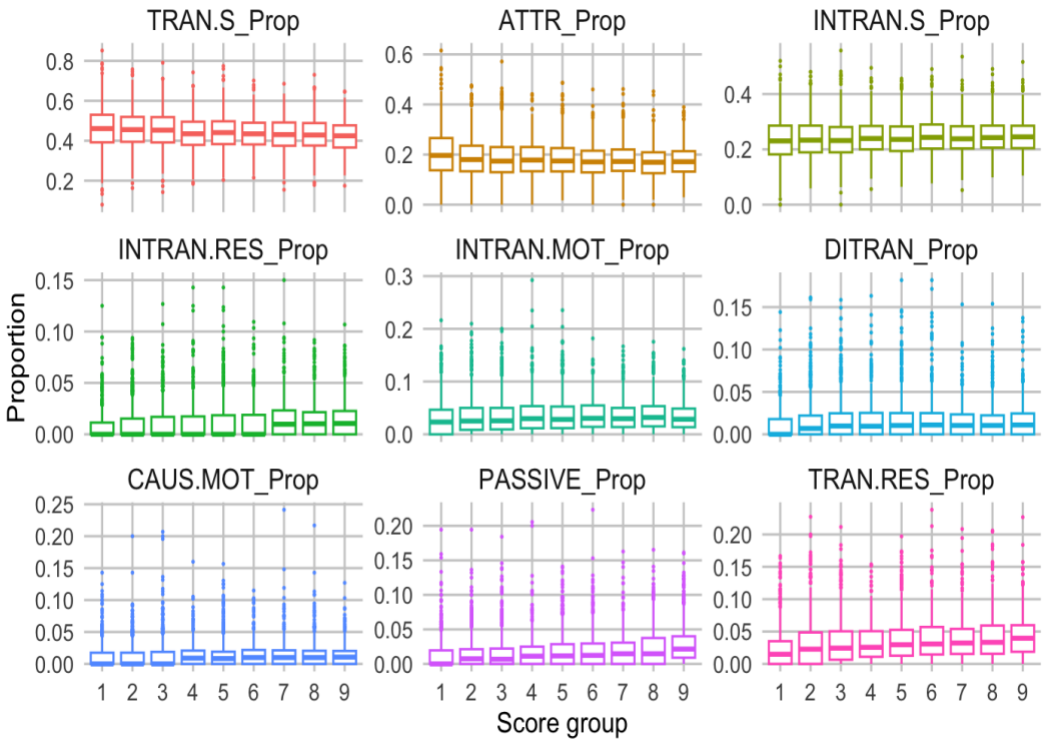

Figure 20. Distribution of proportion indices across proficiency score groups (x-axis) with individual y-axis scales for each index

5.3.1.3 Frequency

More proficient writers were more likely to use low-frequency ASCs and ASC-verb lemma combinations. Constructions that introduce additional semantic arguments (i.e., intransitive resultative, ditransitive, transitive resultative, and caused motion) were relatively rare in the reference corpus and were markedly more frequent in advanced learners' writing (Table 33). This pattern relates to the diversity and proportion results reported earlier and aligns with previous findings on L2 writing and verb-VAC usage (Kyle & Crossley, 2017; Kyle et al., 2021).

Table 33. Frequency of ASC and ASC-verb combinations from the EnCOW-US corpus

	Frequency	Frequent Verbs (Frequency in ASC)
TRAN-S	18,813,622	*have (1,128,293), say (587,160), do (465,033), see (421,297)*
ATTR	7,308,072	*be (6,799,666), seem (166,063), look (79,338), feel (77,463)*
PASSIVE	4,573,146	*use (125,395), make (97,712), call (79,008), do (57,225)*
INTRAN-S	3,991,261	*work (241,923), go (107,853), live (107,509), happen (95,244)*
INTRAN-MOT	1,736,930	*go (253,599), come (205,581), get (95,596), move (59,980)*
CAUS-MOT	1,185,983	*put (87,575), take (60,890), get (59,041), bring (54,715)*
TRAN-RES	1,131,965	*make (229,865), become (198,068), let (115,49)*
DITRAN	549,510	*tell (131095), give (125558), ask (51382), show (19212)*
INTRAN-RES	548,296	*become (198,068), get (71,808), go (35,631), come (35,170)*

Although the correlation between proficiency and the use of low-frequency ASC-verb lemma combinations is modest, it is noticeably stronger than in the parallel spoken-language analysis. Because the two datasets do not track the same learners across modalities, it remains unclear whether this gap reflects individual learner profiles or modality-specific production constraints. Writing generally affords more planning time, lexical precision, and syntactic density than speech (Kim & Crossley 2023); these characteristics may encourage writers to incorporate less-frequent verbs within complex constructions, making the frequency-based ASC-verb lemma combination index especially informative in written contexts. Longitudinal, cross-modal studies that follow the same learners will be essential to determine whether the observed frequency shift represents a stable developmental trajectory or a situational response to the demands of writing.

5.3.1.4 SOA

More proficient writers tended to use verbs that show higher SOA scores when using relatively rare ASCs (i.e., intransitive resultative, passive, and transitive resultative). Of the SOA metrics considered, MI emerged as the most sensitive indicator. As detailed in Section 3.4, MI compares the observed co-occurrence of a verb lemma and a construction with the frequency expected by chance; high values therefore signal a tight, construction-specific bond. For

instance, typical verbs in the transitive resultative construction (e.g., *let, encourage, keep, compel, enable, allow, dissuade*; Table 34) all showed high MI values.

Table 34. SOA (*MI*) of ASC-verb combinations from the EnCOW-US corpus

	verb
TRAN-S	*loath, reweigh, yield, vision, mistake, foreswear, enjoy*
PASSIVE	*detach, suite, threaten, rive, unevolved, affect, network*
INTRAN-S	*aestivate, mother, snore, overflow, condole, occur, sweat*
INTRAN-MOT	*welcome, journey, return, emigrate, inflow, fallback*
CAUS-MOT	*bequeath, endear, wend, betake, arrogate, extricate*
TRAN-RES	*let, encourage, keep, compel, enable, allow, dissuade*
DITRAN	*tell, remind, give, assure, convince, inform, advise*
INTRAN-RES	*become, remain, morph, get, go, degenerate*

Notes. Within each ASC type, verbs are listed from most to least strongly associated.

Notably, these low-frequency constructions are typically anchored by high-frequency, semantically prototypical verbs that are readily retrievable for learners' lexical processing (e.g., TRAN-RES: *let, encourage*; DITRAN: *tell, remind, give*; INTRAN-RES: *become, remain, get*). This pattern suggests that learners may internalize complex argument structures along a continuum that begins with such highly associated verbs; mastering a compact, strongly bonded verb set may therefore facilitate subsequent expansion to a broader range of constructions.

In contrast, the highly frequent transitive-simple construction almost always displays low MI values (≈ 1-2); even its most strongly associated verbs reach only modest MI scores, indicating a generally weak verb-construction bond.[16] The looser verb-construction bond suggests that learners choose the simple grammatical structures mainly because of their ubiquity in the input, rather than any strong verb-specific affinity. The follow-up DeltaStructureCue (DSC) analysis for the transitive-simple construction supports this interpretation. Apart from a small set of verbs that strongly favor the pattern, most verbs display only weak attraction, producing a pronounced long-tail distribution (Figure 21). This diffuse association helps explain

[16] This contrasts sharply with the 4 to 6 range typical of the highly associated verbs of the less frequent ASCs.

DSC's negative correlation with proficiency: as writers advance, they rely less on default, strongly associated verb-construction pairings and increasingly use verbs with weaker constructional ties.

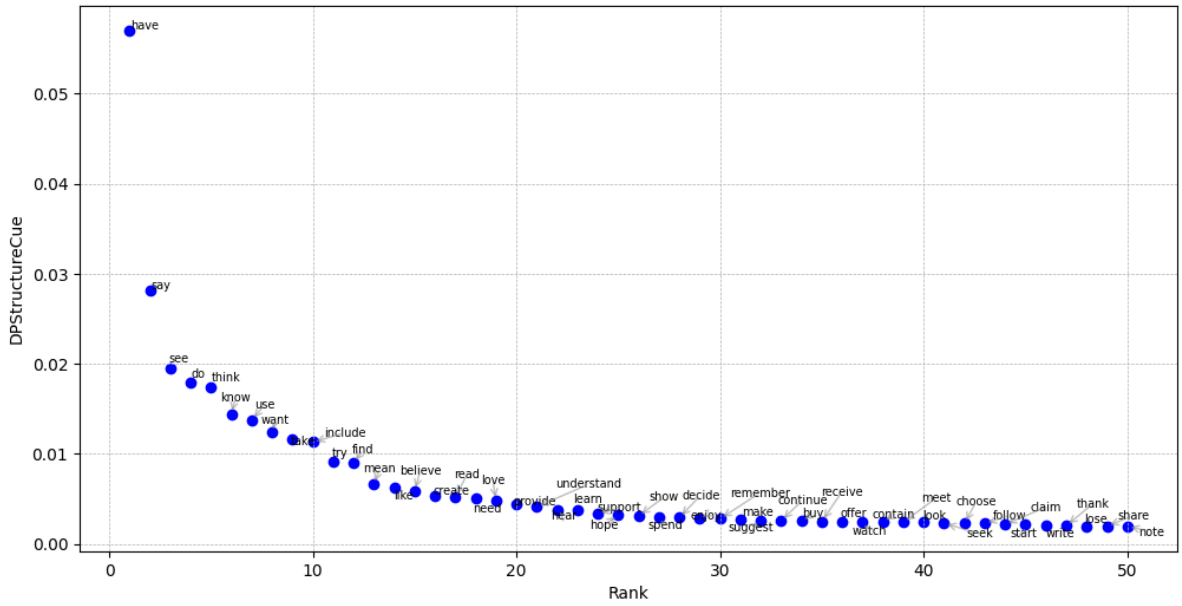

Figure 21. DeltaStructureCue of the top 50 verbs in the transitive simple construction

5.3.2 RQ 2: Extent to which ASC-based indices predict L2 writing proficiency scores

The best multivariate linear regression identified 12 indices that collectively accounted for 14.5% of the score variance, with SOA indices contributing 48.2%, proportion indices 35.36%, and diversity indices 16.42%. Although 14.5 % may be a modest share, it is noteworthy given the multidimensional nature of writing scores and the fact that no other lexical (beyond the verbs), syntactic, or discourse variables were included in the model.

Two interpretation points emerge. First, the large contribution of SOA indices reinforces the idea that lexicalized knowledge (i.e., knowing which verbs are strongly licensed by particular ASCs or vice versa) is a hallmark of advanced writing (Kyle 2016; Kyle & Crossley 2017; Kyle et al. 2021). Second, while proportion and diversity measures are less powerful

individually, they capture distinct yet complementary dimensions of constructional competence. These indices demonstrate the frequency with which specific constructions are used and the breadth of a writer's constructional repertoire (Hwang & Kim, 2023). Together, these three facets (SOA, proportion, diversity) paint a more nuanced picture of lexicogrammatical sophistication and complexity.

Although the 14.5 % figure is lower than the 44.1% variance explained in the earlier spoken corpus, the two results may not be directly comparable. The writing sample was scored on a compressed 1-5 scale (vs. 1-9 for speaking) and came from a different cohort, both of which limit observable variance. In addition, spontaneous speech would likely to rely more on verb-driven, clause-level contructions that the ASC indices capture well, whereas writing allows more planning and heavier use of noun-phrase modification and discourse organization. These register-specific differences may account for the smaller R^2 and point to the need for a follow-up study that analyzes spoken and written output from the same learners under a unified scoring scheme.

5.3.3 RQ 3: Extent to which ASC-based and other indices predict L2 writing proficiency scores

Expanding the analytic scope to include word-, bigram-, and clause-level lexicogrammatical indices substantially strengthened the explanatory power of models predicting L2 writing proficiency. A partial R^2 analysis showed that bigram-related SOA indices explained the largest portion of variance (38.81%), followed by ASC-based indices (18.14%), lexical diversity (14.51%), and lexical sophistication (10.31%), syntactic complexity (10.1%), and lexicogrammatical sophistication (8.07%).

These results highlight the distinctive contribution of ASC-based indices to assessing clausal-level lexicogrammatical complexity and sophistication. Researchers have pointed out that analyses focusing exclusively on syntactic complexity risk overlooking important dimensions of lexicogrammatical sophistication in L2 writing (e.g., Bulté & Housen, 2012; Kyle, 2016; Kyle & Crossley, 2018). Whereas previous studies have relied on either coarse-grained (Lu, 2010, 2011; Hwang et al., 2020; Kim, 2014) or fine-grained syntactic measures (Kyle & Crossley, 2018; Zhang & Lu, 2022), ASC-based indices capture how grammatical structures combine with verb lemmas and semantic argument templates to shape overall text quality, a pattern paralleling findings for VACs (Kyle & Crossley, 2017) and supporting Kyle's (2016) view that syntactic sophistication is central to understanding lexicogrammatical production. The evidence further supports the view that L2 writing proficiency could be better understood through a multifaceted approach (Kyle & Eguchi, 2021); integrating word-, multi-word-, and clause-level indices yields a more nuanced portrait of learners' developmental trajectories.

Future work should investigate the large residual variance by incorporating discourse-level features (e.g., cohesion, rhetorical moves), learner background variables (e.g., L1, educational context), and task characteristics. Examining possible interactions among lexical, ASC-based, and discourse-pragmatic indices may also clarify how different layers of linguistic knowledge jointly contribute to writing proficiency, thereby refining both theoretical models and pedagogical applications.

5.4 Summary

This study aimed to address a gap in the literature by examining L2 writing proficiency through ASC-based indices, moving beyond the field's traditional emphasis on syntactic

complexity. Drawing on the open-source ELLIPSE corpus of 6,482 ESL essays, which were rated on a five-point rubric for syntax, vocabulary, phraseology, and grammar and then averaged into an overall proficiency score, the analyses explored how ASC-based measures relate to proficiency (Section 5.1). Correlation analyses provided a fine-grained view of how different types of ASC index usage are associated with writing proficiency scores (Section 5.2.1). Regression modelling showed that ASC-based indices alone accounted for 14.5 % of the variance in proficiency (Section 5.2.2), and that adding word-, bigram-, and clause-level lexicogrammtical and syntactic features further increased explanatory power (Section 5.2.3). Collectively, these findings suggest that ASC-based indices may capture certain dimensions of clausal-level lexicogrammatical complexity and sophistication from a usage-based perspective. Moreover, they appear to complement conventional lexicogrammatical and syntactic measures, supporting a multivariate approach to analyzing L2 writing proficiency (Section 5.3).

6 Conclusion

This goal of this study was to contribute to the measurement of lexicogrammatical complexity and sophistication in L2 production by developing and evaluating a set of ASC-based indices. To this end, a RoBERTa model was fine-tuned on human-annotated data covering nine ASC types, developing an ASC tagger that automatically labels ASCs in running text. Building on this tagger, an ASC analyzer was created to compute 50 ASC-based indices, including diversity measures, proportional counts, frequency-based scores, and SOA metrics between ASCs and verb lemmas. To assess the contribution of these indices, they were evaluated alongside established lexicogrammatical and syntactic features, with a particular focus on the additional variance in proficiency they could explain. The study addressed three core objectives. First, it examined the correlations between ASC-based indices and human-rated L2 proficiency. Second, it assessed the extent to which ASC-based indices independently predict proficiency scores. Third, it evaluated the additional predictive power contributed by ASC-based indices when combined with other lexicogrammatical and syntactic measures.

6.1 Summary of findings

The study set out to determine how automated ASC-based indices correlate with human-rated L2 proficiency, how well these indices predict proficiency on their own, and the additional predictive value they provide when combined with other lexicogrammatical and syntactic indices. Analyses covered both oral and written modalities to capture a more comprehensive picture of L2 productive proficiency.

6.1.1 Relationship between ASC-based indices and human-rated L2 proficiency

For L2 oral production, ASC diversity indices displayed small but consistent positive correlations with proficiency: more proficient L2 speakers drew on a wider repertoire of constructions. They favored caused-motion, intransitive resultative, passive, and transitive resultative patterns and relied less on attributive, intransitive motion, and transitive simple constructions. They also avoided overusing the most frequent ASCs or ASC-verb lemma combinations, instead selecting less common combinations. Six construction types (i.e., intransitive resultative, passive, caused motion, transitive resultative, transitive simple, intransitive simple) showed moderate-to-small positive correlations with proficiency when paired with their strongly associated verbs, indicating that successful alignment of verbs and constructions is a hallmark of higher-level oral performance.

In written production, a similar, albeit weaker, trend emerged: more proficient L2 writers showed slightly greater ASC diversity and used fewer high-frequency constructions. Specifically, they employed passive, transitive-resultative, and intransitive-resultative patterns more often, while attributive and transitive-simple constructions appeared less frequently. Advanced writers also paired complex constructions (i.e., intransitive resultative, passive, transitive resultative) with their prototypical verbs, suggesting that mastery of a limited set of high-association verbs facilitates the use of sophisticated argument structures. In contrast, the highly versatile transitive-simple pattern showed lower verb-specific associations, reflecting its flexibility.

6.1.2 Extent to which ASC-based indices predict human-rated L2 proficiency

A model containing only ASC-based indices accounted for nearly half of the variance in oral proficiency scores**,** with SOA measures exerting the greatest influence; diversity and proportion indices also contributed, even though many learners still leverage simpler ASCs in their speaking and writing productions. For writing, the ASC-only model explained a smaller yet meaningful share of variance, again led by SOA measures and supported by proportional and diversity indices. Together, these results highlight the central role of ASC-verb relationships in developing both spoken and written productive proficiency.

6.1.3 Extent to which ASC-based and other indices predict human-rated L2 proficiency

Adding ASC-based indices to the established lexicogramamtical and syntactic measures consistently sharpened proficiency predictions. The boost was most pronounced in the speaking domain, where the learner corpus exhibited greater variability and spontaneous constructional choices, allowing ASC information to fill gaps left by other lexicogrammatical metrics. In the writing domain, where exam essays were more standardized, the gains were smaller yet still meaningful: ASC indices revealed clause-level grammatical distinctions that both the lexicogrammatical and syntactic indices overlooked. Taken together, these results support the view that ASCs form an independent lexicogrammatical dimension of complexity and sophistication that complements existing indices across modalities. Because the impact of ASC usage may appear sensitive to register and learner variability, broader empirical studies (drawing on different corpora, tasks, and proficiency ranges) are needed to refine our understanding and further validate these findings.

6.2 Summary of tool and research applications

As part of this study, an open-source NLP was developed to identify and classify ASC types and compute their associated indices. Development began with the creation of an ASC treebank, constructed through a detailed annotation scheme and an iterative curation-annotation-review cycle. The resulting dataset was used to train and evaluate the ASC tagger, which fine-tuned pre-trained language models within a supervised learning framework. Model performance was assessed across both L1 and L2 domains to ensure robustness and generalizability.

The tagger was then integrated into an NLP pipeline, ASC Analyzer, that computes diversity, proportion, frequency, and SOA indices. All materials needed to revisit the dataset, train the model, and consult the ASC annotation guidelines (manual) are publicly available at https://github.com/LCR-ADS-Lab/ASC-Treebank. By releasing the ASC treebank, tagger, and analyzer under an open-source license, the study aims to support researchers seeking an empirical framework for quantifying ASC usage in relation to L2-English productive proficiency assessment and development. In terms of pedagogy, this project may invite further extensions such as automated feedback in L2 speaking or writing assessment. Beyond L2 studies, the resource can support investigations of register and genre variation in English language production, as well as diachronic analyses of constructional change.

6.3 Implications

The findings of the study have four main implications. First, the results support the usage-based constructionist perspective of language learning (Bybee 2010; Diessel 2015; Ellis 2012; Ellis & Ferreira-Junior 2009a, b; Gries & Wulff 2005; Herbst 2016) by showing that ASCs usage systematically correlates with proficiency in both speech and writing. They extend

earlier work on VACs and constructional diversity in L2 development (Kyle 2016; Kyle & Crossley 2017; Kyle et al. 2021; Hwang & Kim 2023; Kim & Ro 2023), supporting the hypothesis that the core ASCs highlighted in construction grammar research (Goldberg 1999) serve as valid markers of grammatical growth in L2 research, especially when investigated in conjunction with verb usage.

Second, ASC-based indices exhibit a strong correspondence with human ratings (particularly in spoken data) thereby addressing a persistent validity gap in automated scoring systems (Enright & Quinlan, 2010; Norris & Ortega, 2009). Although several studies provide empirical evidence that traditional syntactic and grammatical complexity metrics correlate with rater judgments (e.g., Attali, 2013; Burstein & Chodorow, 1999; Chen et al., 2018; Knoch et al., 2014; Powers et al., 2000; Yoon & Bhat, 2012; Zechner et al., 2017), and others (e.g., Cardwell et al., 2024) incorporate these measures into automated scoring systems based on their assumed validity, such broad indices often overlook the clause-level nuances that human raters attend to. Integrating theory-driven, ASC-based measures could provide scoring engines with a finer lens on constructional complexity and sophistication, improving the sensitivity of automated scores. For developers and instructors, ASC-based indices also offer a rapid diagnostic tool for targeted feedback and help ensure that automated assessments more fully capture learners' productive proficiency.

Third, leveraging PLMs to tag nine ASC types overcomes the limitations of earlier rule-based or parser-dependent methods, which often misidentify lexicogrammatically ambiguous constructions (Hwang & Kim, 2023; Kyle, 2016). The success of the fine-tuned RoBERTa tagger demonstrates the potential of large language models for high-precision feature

extraction well beyond ASCs and paves the way for scalable research on other linguistic phenomena that previously challenged human annotation.

Lastly, because, based on the observation, ASC usage reflects distinctions between simpler constructions (e.g., transitive simple) and more complex ones (e.g., caused-motion, transitive resultative), these insights offer English language teachers a concrete basis for teaching core lexicogrammatical constructions. By tracing learners' progression from high-frequency, formulaic patterns to more varied and information-dense structures, ASC-based analysis can inform syllabus design (Pienemann, 1985; Hinkel, 2003) and support the development of materials that scaffold the use of sophisticated grammar and lexis (Hamp-Lyons & Henning, 1991; Nation, 2001), ultimately promoting greater lexicogrammatical richness in learner production.

6.4 Limitations

This study is subject to several limitations. First, while the tool leveraged PLMs to automatically extract ASCs and related indices, it may have introduced potential errors or biases stemming from the model's training data and/or underlying algorithms. These biases could have affected the accuracy and reliability of the computed indices. To address this limitation from an annotation perspective, future work could refine the annotated datasets used for model fine-tuning, incorporate linguistically diverse training data, and rigorously evaluate tool performance using out-of-domain texts.

Second, the learner corpora employed in this study may not have captured the full spectrum of linguistic backgrounds or proficiency levels typically observed in the L2 population. Consequently, the external validity of the findings could be constrained. Future work should

incorporate more diverse learner groups, varying in first-language typology, exposure contexts, and proficiency bands, and adopt hierarchical or mixed-effects frameworks that model inter- and intra-group variability in ASC usage. In addition, the tasks employed in this study might not have reflected authentic language use. Future research could broaden generalizability by examining a wider variety of task types (e.g., narrative, argumentative, interactive dialogues). Moreover, directly comparing linguistic performance across multiple modalities (e.g., spoken versus written outputs from the same L2 learner) could yield insights into modality-specific lexicogrammatical complexity and sophistication features.

Third, the analysis focused on a limited set of ASCs, which may obscure subtler layers of lexicogrammatical development. More fine-grained analysis could be helpful in future work. For example, identifying major constructions such as simple transitive and further categorizing them according to the grammatical structure of their object arguments (e.g., noun phrases, finite complements, infinitival complements) would offer a more systematic and nuanced depiction of learners' lexicogrammatical complexity and sophistication.

Lastly, although proficiency ratings in this study adhered to standardized procedures, inherent subjectivity remains a concern. Employing multiple raters or triangulating these ratings with standardized test scores could help mitigate potential biases, thus enhancing the validity and reliability of proficiency assessments. Importantly, additional empirical studies are needed to validate and build upon these findings, thereby deepening our understanding of lexicogrammatical complexity and sophistication as it relates to L2 productive proficiency.

References

Abdi Tabari, M., Khezrlou, S., & Tian, Y. (2024). Verb argument construction complexity indices and L2 written production: Effects of task complexity and task repetition. *Innovation in Language Learning and Teaching*, *18*(1), 1-16.

ACTFL-ALC Press. (1996). Standard Speaking Test manual.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., & Zhu, H. (2015). Generating high quality proposition banks for multilingual semantic role labeling. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 397-407). Association for Computational Linguistics.

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, *1*(1), 96-129.

Allen, V. F. (1983). *Techniques in teaching vocabulary*. Oxford: Oxford University Press.

Andersen, Ø. E., Nioche, J., Briscoe, T., & Carroll, J. A. (2008). The BNC parsed with RASP4UIMA. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)* (pp. 865-869). European Language Resources Association.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In *Handbook of automated essay evaluation* (pp. 181-198). Routledge.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, *26*(2), 390-395.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, *11*(1), 17-34.

Barton, K. (2023). *MuMIn: Multi-Model Inference* (R package version 1.47.1). Retrieved from https://CRAN.R-project.org/package=MuMIn

Bencini, G. M., & Goldberg, A. E. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, *43*(4), 640-651.

Berger, C., Crossley, S. & Kyle, K. (2017). Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second-Language Production. *Applied Linguistics*, 40(1), 22–42.

Berger, C., Crossley, S., & Skalicky, S. (2019). Using lexical features to investigate second language lexical decision performance. *Studies in Second Language Acquisition*, *41*(5), 911-935.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016, August). Universal Dependencies for Learner English. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 737–746). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1070

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, *45*(1), 5-35.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639-668.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman,

Bonial, C. N. (2014). *Take a look at this! Form, function and productivity of English light verb constructions* [Doctoral dissertation, University of Colorado at Boulder]. ProQuest Dissertations & Theses Global.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, *32*, 21.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

Bulté, B., Housen, A., & Pallotti, G. (2024). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*.

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, *91*, 102246.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 1-8). Association for Computational Linguistics.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991-997.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904-911.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Cardwell, R., Naismith, B., LaFlair, G. T., & Settles, B. (2024). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Casal, J. E., Shirai, Y., & Lu, X. (2022). English verb-argument construction profiles in a specialized academic corpus: Variation by genre and discipline. *English for Specific Purposes*, *66*, 94-107.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740–750). Association for Computational Linguistics.

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRater℠ v. 5.0 engine* (Research Report No. RR–18–10). Educational Testing Service. https://doi.org/10.1002/ets2.12198

Chen, H., Zhang, M., Li, J., Zhang, M., Øvrelid, L., Hajič, J., & Fei, H. (2025). *Semantic role labeling: A systematical survey* (arXiv Preprint No. arXiv:2502.08660). arXiv. https://arxiv.org/abs/2502.08660

Choi, J., & Sung, M. C. (2020). Utterance-based measurement of L2 fluency in speaking interactions: A constructionist approach. *English Teaching*, *75*(1), 105-126.

Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition*, *41*(1-3), 83-121.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In L. R. Gleitman & B. Landau (Eds.), *The acquisition of the lexicon* (pp. 1-33). Psychology Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307-334.

Crossley, S., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580.

Crossley, S., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., ... & Boser, U. (2023). The english language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, *9*(2), 248-269.

Davies, M. (2020). *The Corpus of Contemporary American English (COCA)*. Brigham Young University. https://www.english-corpora.org/coca/

de Marneffe, M. C., & Manning, C. D. (2008). *Stanford typed dependencies manual* (Technical report). Stanford University. https://nlp.stanford.edu/software/dependencies_manual.pdf

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255-308.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Diessel, H. (2004). *The acquisition of complex sentences* (Vol. 105). Cambridge University Press.

Diessel, H. (2015). Usage-based construction grammar. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 296–322). De Gruyter Mouton.

Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing [Preprint]. *arXiv*. https://arxiv.org/abs/1611.01734

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381-400.

Eguchi, M., & Kyle, K. (2023). L2 collocation profiles and their relationship with vocabulary proficiency: A learner corpus approach. *Journal of Second Language Writing*, *60*, 100975.

Eguchi, M., & Kyle, K. (2023). Span identification of epistemic stance-taking in academic written English. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 429–442). Association for Computational Linguistics.

Ellis, N. C. (2002a). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 297-339.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*(4), 559-617.

Ellis, N. C., & Ferreira-Junior, F. (2009a). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188-221.

Ellis, N. C., & Ferreira-Junior, F. (2009b). Construction learning as a function of frequency, frequency distribution, and function. *The Modern language journal*, *93*(3), 370-385.

Ellis, N. C., & Wulff, S. (2014). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (2nd ed., pp. 87–105). Routledge.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317-334.

Ettinger, A., Hwang, J. D., Pyatkin, V., Bhagavatula, C., & Choi, Y. (2023). "You Are an Expert Linguistic Annotator": Limits of LLMs as Analyzers of Abstract Meaning Representation**.** In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8250–8263). Association for Computational Linguistics.

Faerch, C., Haastrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Multilingual Matters.

Fillmore, C. J. (1968). Lexical entries for verbs. *Foundations of Language, 4*, 373–393.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, *280*(1), 20-32.

Fillmore, C. J. (1979). On fluency. In D. Kempler & W. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85-102). New York: Academic Press.

Fillmore, C. J. (1988). The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society* (pp. 35-55).

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. (2003). Background to framenet. *International Journal of Lexicography*, *16*(3), 235-250.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, 501-538.

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176-187.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. In A. Kumar, O. Levy, & Y. Miyao (Eds.), *Proceedings of the Workshop for NLP Open Source Software (NLP-OSS 2018)* (pp. 1–6).

Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: An introductory course*. Routledge.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013, October). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Miller, K. I. Martin, C. M. Eddington, & A. Henery (Eds.), *Proceedings of the 31st Second Language Research Forum* (pp. 240–254). Cascadilla Proceedings Project.

Gerdes, K., & Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. In A. Friedrich & K. Tomanek (Eds.), *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 131–140). Association for Computational Linguistics.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, *28*(3), 245-288.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2013). Constructionist Approaches. *Oxford Handbooks Online*.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3).

Goldberg, A. E., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 468-477.

Gilquin, G. (2019). Light verb constructions in spoken L2 English: An exploratory cross-sectional study. *International Journal of Learner Corpus Research*, *5*(2), 181-206.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*(S1), 228-255.

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions?. *Annual Review of Cognitive Linguistics*, *3*(1), 182-200.

Grömping, U. (2023). *relaimpo: Relative importance of regressors in linear models* (Version 2.2.1) [R package]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=relaimpo

Guiraud, P. (1960). *Probl`emes et m´ethodes de la statistique linguistique[Problems and methods of linguistic statistics]*. Dordrecht: Reidel.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41*(3), 337-373.

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*, e5.

Herbst, T. (2016). Foreign language learning is construction learning–what else? Moving towards Pedagogical Construction Grammar. *Applied Construction Grammar*, *32*, 56-96.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4129–4138). Association for Computational Linguistics.

Hinkel, E. L. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *Tesol Quarterly*, *37*(2), 275-301.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python* (Version 2.0) [Computer software]. Explosion AI. https://spacy.io

Housen, A. (2008). A corpus-based study of the L2-acquisition of the English verb system. In *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77-116). John Benjamins Publishing Company.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1-20). John Benjamins Publishing Company.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). National Council of Teachers of English.

Hwang, H., Jung, H., & Kim, H. (2020). Effects of written versus spoken production modalities on syntactic complexity measures in beginning-level child EFL learners. *The Modern Language Journal*, *104*(1), 267-283.

Hwang, H., & Kim, H. (2023). Automatic analysis of constructional diversity as a predictor of EFL students' writing proficiency. *Applied Linguistics*, *44*(1), 127-147.

Ishikawa, S. I. (2011). A new horizon in learner corpus studies: The aim of the ICNALE Project. In G. Weir, S. I. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp. 3–11). University of Strathclyde Press.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, *12*(2), 119-125.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017*). An introduction to statistical learning: With applications in R*. Springer.

Johnson, W. (1944). Studies in language behavior 1: A program of research. *Psychological Monographs, 56*, 1-15.

Kameen, P. T. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 343–364). Washington, D.C.: TESOL.

Kim, H., Shin, G. H., & Sung, M. C. (2023). Constructional complexity as a predictor of Korean EFL learners' writing proficiency. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *36*(2), 436-466.

Kim, H., & Rah, Y. (2016). Effects of verb semantics and proficiency in second language use of constructional knowledge. *The Modern Language Journal*, *100*(3), 716-731.

Kim, H., & Rah, Y. (2019). Constructional processing in a second language: The role of constructional knowledge in verb-construction integration. *Language Learning*, *69*(4), 1022-1056.

Kim, H., & Ro, E. (2023). Assessment of sentence sophistication in L2 spoken production: Expansion of verbs and argument structure constructions. *System*, *119*, 103175.

Kim, H., & Ro, E. (2024). Usage-based approaches to assessing syntactic sophistication in second language writing: Interaction of genre and proficiency. *Journal of Second Language Writing*, *65*, 101131.

Kim, J. Y. (2014). Predicting L2 Writing Proficiency Using Linguistic Complexity Measures: A Corpus-Based Study. *English Teaching*, *69*(4).

Kim, M., & Crossley, S. (2023). Lexical and phraseological differences between second language written and spoken opinion responses. *Frontiers in Psychology*, *14*, 1068685.

Kim, M., Crossley, S., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120-141.

Kim, S., Ko, H., & Yang, H. K. (2020). Telicity and mode of merge in L2 acquisition of resultatives. *Language Acquisition*, *27*(2), 117-159.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. In *Psychology of learning and motivation* (Vol. 7, pp. 1-41). Academic Press.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test. *ETS Research Report Series*, *2014*(2), 1-74.

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(1), 55–73.

Koizumi, R., & Hirai, A. (2012). Comparing the story retelling speaking test with other speaking tests. *JALT Journal*, *34*, 35–60.

Kulmizev, A., Ravishankar, V., Abdou, M., & Nivre, J. (2020). Do neural language models show preferences for syntactic formalisms? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4077–4091). Association for Computational Linguistics.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine-grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. ProQuest Dissertations & Theses Global.

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), 757-786.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513-535.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030-1046.

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*(4), 781-812.

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 126–151). Multilingual Matters.

Kyle, K., & Eguchi, M. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal*, *107*(2), 531-564.

Kyle, K., Eguchi, M., Miller, A., & Sither, T. (2022). A dependency treebank of spoken second language English. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 39–45). Association for Computational Linguistics.

Kyle, K., & Sung, H. (2023). An argument structure construction treebank. In C. Bonial & H. Tayyar Madabushi (Eds.), *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)* (pp. 51–62). Association for Computational Linguistics.

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, *46*(1), 278-299.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume I: Theoretical prerequisites* (Vol. 1). Stanford university press.

Langacker, R. W. (1990). *Concept, image, and symbol: The cognitive basis of grammar* (Vol. 1). Walter de Gruyter.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579-589.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Li, H. (2015). Relationship between measures of syntactic complexity and judgments of EFL writing quality. In *Proceedings of 2015 youth academic forum on linguistics, literature, translation and culture* (pp. 216-222). American Scholars Press.

Li, H., & Yu, X. (2024). Verb argument constructions in argumentative essays by college-level Asian learners of English: Exploring the effects of English proficiency, acquisition context, and topic. *Journal of Second Language Writing*, *65*, 101127.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis* (Vol. 4). Glenview, IL: Scott, Foresman.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. LiberFörlag.

Liu, Y., & Lu, X. (2024). Development of verb argument constructions in L2 English learners: A close replication of research question 3 in Römer and Berger (2019). *Studies in Second Language Acquisition*, 1-19.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach [Preprint]. *arXiv*. https://arxiv.org/abs/1907.11692

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Meara, P., & Bell, H. (2001). P-Lex: A Simple and Effective Way of Describing the lexical Characteristics of Short L2 Tests. *Prospect*, *16*(3), 5-19.

Menard, S. (2001). *Applied logistic regression analysis*. SAGE publications.

Miaschi, A., & Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: An in-depth linguistic investigation. In S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, & H. Hajishirzi (Eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 110–119). Association for Computational Linguistics.

Mostafa, T., & Crossley, S. (2020). Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts. *Journal of Second Language Writing*, *49*, 100730.

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition*, *21*(1), 49-80.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language* (Vol. 10, pp. 126-132). Cambridge: Cambridge University Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, *26*(3), 619-653.

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., … & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659–1666). European Language Resources Association.

Nivre, J., de Marneffe, M. C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection [Preprint]. *arXiv*. https://arxiv.org/abs/2004.10643

Nivre, J., De Smedt, K., & Volk, M. (2004). Treebanking in Northern Europe: A white paper. *Nordisk Sprogteknologi, 2000–2004*, 169–182.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578.

O'Donnell, M. B., & Ellis, N. C. (2010). Towards an inventory of English verb argument constructions. In M. Sahlgren & O. Knutsson (Eds.), *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 9–16). Association for Computational Linguistics.

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.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71-106.

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.

Park, J. H., & Sung, M. C. (2024). Expansion of verb-argument construction repertoires in L2 English writing. *International Review of Applied Linguistics in Language Teaching*, *62*(2), 903-925.

Perdue, C. (1993). *Adult language acquisition: Crosslinguistic perspectives*. Cambridge: Cambridge University Press.

Pienemann, M. (1985). Learnability and syllabus construction. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 23–75). Multilingual Matters.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning*, *47*(1), 101-143.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *ETS Research Report Series*, *2000*(2), i-23.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Read, J. (2000). *Assessing vocabulary*. Cambridge university press.

Rhee, S. C., & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *The Journal of the Korea Contents Association*, *14*(11), 1019-1029.

Romain, L. (2022). Putting the argument back into argument structure constructions. *Cognitive Linguistics*, *33*(1), 35-64.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(1), 140–162.

Römer, U., & Berger, C. M. (2019). Observing the emergence of constructional knowledge: Verb patterns in German and Spanish learners of English at different proficiency levels. *Studies in Second Language Acquisition*, *41*(5), 1089-1110.

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2014). Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology. *The Modern Language Journal*, *98*(4), 952-975.

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.

Saito, K. (2020). Multi-or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548-588.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, *38*(4), 677-701.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* (pp. 28–34). Institut für Deutsche Sprache.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). European Language Resources Association.

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Shi, P., & Lin, J. (2019). Simple BERT models for relation extraction and semantic role labeling [Preprint]. *arXiv*. https://arxiv.org/abs/1904.05255

Silveira, N., Dozat, T., de Marneffe, M. C., Bowman, S., Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)* (pp. 2897–2904). European Language Resources Association.

Sung, H. (2019). Korean EFL Learners' Processing of English Caused-Motion Construction. *English Teaching*, *74*(1), 49-73.

Sung, M. C., & Kim, H. (2022). Effects of verb–construction association on second language constructional generalizations in production and comprehension. *Second Language Research*, *38*(2), 233-257.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description*, *3*(99), 36-149.

Tan, M. Y. J., & Biswas, R. (2012). The reliability of the Akaike information criterion method in cosmological model selection. *Monthly Notices of the Royal Astronomical Society, 419*(4), 3292-3303.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

van der Laken, P., & Lambert, L. (2023). *corrtable: Creates and saves out a correlation table with significance levels indicated* (Version 0.1.1) [R package]. https://cran.r-project.org/package=corrtable

Van Nguyen, M., Lai, V. D., Veyseh, A. P. B., & Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 80–90). Association for Computational Linguistics.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy. *& Complexity. Hawaii: University of Hawaii*.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53-67.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.

Viberg, Å. (2002). Basic verbs in second language acquisition. *Revue française de linguistique appliquée*, (2), 61-79.

Yoon, S. Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. In J. Tsujii, J. Henderson, & M. Paşca (Eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 600–608). Association for Computational Linguistics.

Yoon, S. Y., Lu, X., & Zechner, K. (2019). Features measuring vocabulary and grammar. In K. Evanini, X. Wang, & A. Loukina (Eds.), *Automated speaking assessment* (pp. 123–137). Routledge.

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534-561.

Zechner, K., Yoon, S. Y., Bhat, S., & Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior*, *76*, 672-682.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505.

Zhang, X., & Lu, X. (2022). Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing*, *51*, 100597.

Appendix

A. Hyperparameter settings for the transformer-based NER model

Hyperparameter	Selected Value
Num hidden units	200
Embedded vector space	50
Number of layers	1
Dropout rate of layers	0.5
Beam size	1
Attention type	soft
Optimization algorithm	Adam
Learning rate	0.001
Num epochs	60
Batch size (training)	50
Max grad norm	5.0
GPU allocator	pytorch
Seed	0
Batch size (inference)	128
Hidden width	64
Maxout pieces	2
Max batch items	4096
Transformer model name	roberta-base
Window size (get spans)	128
Stride (get spans)	96
Accumulate gradient	3
Dropout (training)	0.1
Patience	1600
Max epochs (early stopping)	0
Max steps	20000
Eval frequency	200
Initial learning rate (scheduler)	0.00005
Warmup steps	250
Total steps (learning rate schedule)	20000
L2 regularization	0.01
Grad clip	1.0
Epsilon (optimizer)	0.00000001

B. JLE scoring rubric

Score	Scoring descriptions
9	A Level-9 speaker can proficiently respond to any topics ranging from familiar ones to those of general interest. He/she can comfortably speak in any tense, for example, to narrate and describe and can effectively deal with unexpected complications as well. In addition, a speaker at this level can construct his/her response in a logical paragraph-like structure. Though few unconsciously made minor errors in grammar and word choices may be present, such do not impede comprehension at all.
8	A Level-8 speaker can proficiently respond to various topics ranging from familiar ones to those of general interest. He/she is able to deal with unexpected complications most of the time. Though rare, flaws in grammar are still present. Tense control may still weaken in certain cases, and the speaker may have some difficulty in complex sentence construction. The responses are mostly organized but sometimes lack fluency and/or may include minor word choice errors; needless to say, they do not have a significant impact on listeners' comprehension.
7	A Level-7 speaker can communicate with proficiency necessary to live and survive in English-speaking countries. He/she is able to deal with complicated situations as well, but effort is required in doing so as grammar/fluency control and speech organization may weaken. Nonetheless, a speaker at this level has noticeable strengths supporting their proficiency such as abundant volume or native-like pronunciation.
6	A Level-6 speaker can communicate with proficiency necessary to live and survive in English-speaking countries. The speaker can somewhat effortlessly string simple sentences together to express his/her thoughts; however, as the sentences become longer and more complex, fluency and grammar control sometimes weaken. Tense control errors may still often be present. Pronunciation varies from speaker to speaker. Some may sound native-like whereas others are still influenced by their native language.
5	A Level-5 speaker can maintain simple communication by talking about familiar topics, answering and asking simple questions. The speaker can also add extra information and details to his/her responses, but as sentences become longer and more complex, accuracy weakens. For example, the speaker's grammar control and fluency may weaken, and/or it may require much time for the speaker to complete them. Word choices and pronunciation are still influenced by the speaker's native language; however, listeners used to non-native English speakers would not have trouble understanding the responses.
4	A Level-4 speaker can maintain simple communication by talking about familiar topics and asking simple questions. A speaker at this level can connect simple short sentences to convey his/her thoughts, but fluency is disturbed doing so. With effort, the speaker can manage to respond to what has been asked, but he/she still cannot actively interact. The speaker's pronunciation and word choices may still be influenced by his/her native language, but the impact is insignificant, and listeners used to non-native English speakers would not have trouble understanding him/her.
3	In addition to memorized set phrases, a Level-3 speaker, at times, creates simple short sentences to convey his/her thoughts. However, the speaker is only able to do so when the content of the response is very familiar to him/her, and major errors in grammar and word choices impeding comprehension are still present. Since a great amount of effort is required to create, the responses are often slow, thus requiring listeners' patience. In addition, the pronunciation of a speaker at this level is still influenced by his/her native language and is, at times, difficult to understand without clarification.
2	With a great amount of effort, a Level-2 speaker may provide the bare minimum information necessary to maintain communication when answering simple questions regarding his/her everyday life. However, the responses are mainly just a combination of words, phrases, and memorized set expressions. There are long pauses in the responses, and in some cases, we may hear the speaker simply repeat what was heard in the question. The speaker may attempt to create in sentences; however, major errors in grammar and word choices are frequent. Even listeners who are used to hearing non-native English speakers have difficulty understanding a speaker at this level.
1	A Level-1 speaker cannot communicate in English. The speaker may identify him/herself and make simple greetings using memorized phrases. However, in most cases, the speaker can only speak in fragments of sentences, basically just listing simple vocabulary such as numbers, days of the week, colors, and so on. He/she can rarely respond to questions, and even when showing some sort of response, it takes a tremendous amount of time doing so. In addition, the pronunciation of a speaker at this level is heavily influenced by his/her native language making it significantly difficult to understand the response.

Notes. The original rubric was accessed at http://tsst.alc.co.jp/sst/e/index.html (January 2020). However, the corresponding url is no longer accessible due to the changes made by the ALC press, as noted in the study by Kyle et al. (2024).

C. ELLIPSE scoring rubric

	Holistic	Analytic					
	Overall	**Cohesion**	**Syntax**	**Vocabulary**	**Phraseology**	**Grammar**	**Conventions**
5	Native-like facility in the use of language with syntactic variety; appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.	Text organization consistently well controlled using a variety of effective linguistic features (e.g., reference and transitional words/phrases) to connect ideas across sentences and paragraphs; appropriate overlap of ideas.	Flexible and effective use of a full range of syntactic structures (simple, compound, complex); rare minor and negligible errors in sentence formation.	Wide range of vocabulary flexibly and effectively used to convey precise meanings; skillful use of topic-related terms and less common words; rare negligible inaccuracies in word use.	Flexible and effective use of a variety of phrases (e.g., idioms, collocations, lexical bundles) to convey precise and subtle meanings; rare minor inaccuracies that are negligible.	Command of grammar and usage with few or no errors.	Consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors nonexistent or negligible.
4	Facility in the use of language with syntactic variety and a range of words/phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication.	Organization generally well controlled; a range of cohesive devices (e.g., reference, transitional words/phrases) used appropriately to connect ideas; generally appropriate overlap of ideas.	Appropriate use of a variety of syntactic structures (simple, compound, complex); occasional errors or inappropriateness in sentence formation.	Sufficient range of vocabulary to allow flexibility and precision; appropriate use of topic-related terms and less common lexical items.	Appropriate use of a variety of phrases (e.g., idioms, collocations, lexical bundles); occasional inaccuracies or colloquialisms.	Minimal errors in grammar and usage.	Generally consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors are few and not distracting.
3	Facility limited to the use of common structures and generic vocabulary; organization generally controlled, though connections may be absent or unsuccessful; errors in grammar, syntax, and usage. Communication is impeded by language inaccuracies in some cases.	Organization generally controlled; cohesive devices used but limited in type; some repetitive, mechanical, or faulty use of cohesion within and/or between sentences and paragraphs.	Simple, compound, and complex syntactic structures present, though the range may be limited; some errors in sentence formation, especially in more complex sentences.	Minimally adequate range of vocabulary for the topic; imprecise use of subtle word meanings; topic-related terms used only occasionally; attempts to use less common vocabulary but with some inaccuracy.	Evident use of phrases (e.g., idioms, collocations, lexical bundles) but without much variety; some noticeable repetitions and misuses.	Some errors in grammar and usage.	Developing use of conventions to convey meaning; errors in spelling, capitalization, and punctuation that are sometimes distracting.
2	Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or	Organization only partially developed with a lack of logical sequencing of ideas; some basic cohesive devices used but with	Some sentence variation is used; many sentence structure problems.	Narrow range of vocabulary to convey basic and elementary meaning; topic-related terms used inappropriately; errors in word formation and	Narrow range of phrases (e.g., collocations, lexical bundles) used to convey basic and elementary meaning; many repetitions	Many errors in grammar and usage.	Variable use of conventions; spelling, capitalization, and punctuation errors are frequent and distracting.

	unsuccessful. Communication is impeded in many instances by language inaccuracies.	inaccuracy or repetition.		word choice that may distort meanings.	and/or misuses of phrases.		
1	A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication is impeded in most cases by language inaccuracies.	No clear control of organization; cohesive devices not present or unsuccessfully used; presentation of ideas is unclear.	Pervasive and basic errors in sentence structure and word order that cause confusion; basic sentence errors are common.	Limited vocabulary often inappropriately used; limited control of word choice and word forms; little attempt to use topic-related terms.	Memorized chunks of language or simple phrasal patterns predominate; many repetitions and misuses of phrases.	Errors in grammar and usage throughout.	Minimal use of conventions; spelling, capitalization, and punctuation errors throughout.

Notes. The original rubric was accessed at https://github.com/scrosseye/ELLIPSE-Corpus (February 2025). The yellow-shaded columns indicate the categories that were averaged to compute the L2 writing-proficiency score used in the Chapter 5 study.