AUTOMATIC ANALYSIS OF EPISTEMIC STANCE-TAKING IN ACADEMIC ENGLISH WRITING:

A SYSTEMIC FUNCTIONAL APPROACH

by

MASAKI EGUCHI

A DISSERTATION

Presented to the Department of Linguistics
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2023

DISSERTATION APPROVAL PAGE

Student: Masaki Eguchi

Title: Automatic Analysis of Epistemic Stance-Taking in Academic English Writing: A Systemic Functional Approach

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Linguistics by:

Kristopher Kyle	Chairperson
Julie M. Sykes	Core Member
Vsevolod M. Kapatsinski	Core Member
Thien Huu Nguyen	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2023

© 2023 Masaki Eguchi
This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike (United States) License.

DISSERTATION ABSTRACT

Masaki Eguchi

Doctor of Philosophy

Department of Linguistics

June 2023

Title: Automatic Analysis of Epistemic Stance-Taking in Academic English Writing: A Systemic Functional Approach

Existing linguistic textual measures that investigate features of academic writing often focus on lexis, syntax, and cohesion, despite writing skills being considered more complex and multifaceted (e.g., Sparks et al., 2014). For this reason, writing assessment researchers seek ways to measure and assess various textual features beyond the traditional ones, including discourse moves and steps (Cotos, 2014), source use (Burstein et al., 2018; Kyle, 2020), and essay argument structures (Fiacco et al., 2022). The present dissertation attempts to extend this research by proposing an automated analysis of rhetorical discourse features of epistemic stance-taking strategies.

Drawing on a theoretical framework of the engagement system from Appraisal Analysis (Martin & White, 2005), which originates from the Sydney School of the systemic functional discourse analysis tradition, the dissertation develops and evaluates a series of end-to-end machine learning models to conduct automated engagement resource analysis. The experiment in Study 1 indicated that the developed system can perform as well as (or even outperform) trained annotators' intercoder agreement. Study 2 uses the natural language processing (NLP) systems to conduct the first large-scale analysis of engagement resources in university written assignments across genres and disciplines. The findings suggested that the registers of university writings are

far more complex and nuanced than simple characterization by genres or disciplines.

Study 3 investigates whether the developed measures of rhetorical features of engagement can provide additional information above and beyond the traditional linguistic measures at the levels of lexis, syntax, and cohesion, for modeling professional ratings of essay qualities in a standardized second language proficiency assessment. The results indicate that the features of engagement (particularly the diversity of rhetorical strategies) can complement the existing measures in predicting essay quality.

The three studies together indicate that the proposed machine-learning approach is beneficial to scale up the analysis of rhetorical discourse features in academic writing for research and educational purposes. The dissertation concludes with a call for increased collaboration among discourse analysts, second language researchers, assessment researchers, and computational linguists to define essential textual features for writing assessments across contexts and automate the analysis of such constructs (Lu, 2021, Burstein et al., 2016).

ACKNOWLEDGMENTS

This Ph.D. could not have been possible or even started without the unwavering guidance, support, and encouragement of many individuals. First and foremost, I am truly fortunate to have Dr. Kristopher Kyle as my mentor throughout my Ph.D. journey. Since the very first time I reached out to him as an MA student interested in TAALES (almost seven years ago), Kris has always been caring and supportive. I learned so much from him about learner-corpus research, NLP, language assessment, responding to reviewers, and navigating the job market. Through these instances, he has shown me what it means to be an exceptional mentor. Kris has always made himself available to me, and I cannot simply count how many times I have said "thank you" to him over the years. His wholehearted and constant encouragement helped me overcome multiple obstacles I faced during my Ph.D. Without Kris' guidance and support, my Ph.D. project would have never achieved its current state. My deepest, sincerest gratitude goes to Kris.

I wish to express my sincere appreciation to the other members of my committee: Professor Julie Sykes, Professor Vsevolod (Volya) Kapatsinski, and Professor Thien Huu Nguyen. Since I came to the University of Oregon, Julie has encouraged and helped me at various stages of my Ph.D. journey. Her passion for pragmatics inspired me to work on the interface between the functional approach to language and automated writing assessment in my dissertation. Volya has also inspired me regarding the sophisticated quantitative methodology and his expertise in psycholinguistic approaches to lexical processing. He is also one of the most approachable faculty members in the department, and I appreciate the great discussion we had at BFLs and other department gatherings. I am grateful and lucky to have Dr. Thien Huu Nguyen as the institutional representative on my committee. He kindly welcomed me, a linguistic Ph.D.

student with minimal experience, to his Ph.D. level seminar on Natural Language Processing in the Computer Science Department. I gained so much inspiration from the sessions, which were helpful in completing my dissertation project. Also, I would like to thank Dr. Daniel Anderson (former additional Core Member). His passion for Bayesian statistics convinced me into the Bayesian world. I am also grateful to the other faculty members of the Linguistics Department, Language Teaching Studies (LTS), and the Center for Applied Second Language Studies (CASLS) at the University of Oregon for supporting my Ph.D. for the past four years.

Next, I would like to express my sincere gratitude to Aaron Miller and Ryan Walker (in alphabetical order), who dedicated countless hours and put in painstaking efforts to produce the best possible annotations. Their meticulous work was so critical that, without their help, I could not have even started Study 1 of the dissertation, let alone complete all three projects in a timely fashion.

I would also like to extend my appreciation to those who shaped my academic interests. First, I would like to express my sincere gratitude to my BA and MA mentor, Professor Tetsuo Harada, who equipped me with a bird's-eye view of the field of applied linguistics and possible dots to connect. Special thanks go to Professor Kazuya Saito, who encouraged me to start graduate study in the first place and continue it even during challenging times. I also thank Professor Yasuyo Sawaki, who shaped how I approach topics in language assessment and taught me the foundation of statistical analyses. I would also like to thank the following individuals who helped me in various facets of the academic journey: Professors Stuart Webb, Yuichi Suzuki, and Ronald Heck.

I am also indebted to many extraordinary people from the University of Oregon, the University of Hawai‘i at Mānoa, and other parts of the world, who have always been inspiring

and supportive (in alphabetical order): Ivan Bondoc, Ann Tai (Diane) Choe, Carla Consolini, Carissa Diantoro, Kurtis Foster, Ksenia Gordeeva, Dan Holden, Ryo Maie, Thu Hà Nguyễn, Sayako Nakamura, Hitoshi Nishizawa, Mery Díez-Ortega, Jill Potratz, Milntra (Min) Raksachat, Kiyotaka Suga, Shungo Suzuki, Aki Tsunemoto, Takumi Uchihara, Fred Zenker, and other members of the LING community at UO, the SLS community at UH, and Japan Second Language Acquisition Research Forum.

My Ph.D. study was only possible with the financial support I received at the University of Hawai‘i and the University of Oregon. I wish to express my sincere gratitude to the Crown Prince Akihito Scholarship Foundation and its board of trustees for supporting the first year of my Ph.D. study at the University of Hawai‘i at Mānoa. I would also like to thank the Japan Student Service Organization (JASSO) scholarship for supporting the last three years of my Ph.D. study at the University of Oregon.

The current dissertation project was either directly or indirectly supported by the following sources of funding: the Duolingo English Test's Doctoral Dissertation Award 2022, the International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant 2022, the National Federation of Modern Language Teachers Association and the Modern Language Journal (NFMLTA-MLJ) Dissertation Writing Support Grant 2022, Mary Spaan Research Grant 2020, the Graduate Student Research Award at the Linguistics Department, University of Oregon, and Dr. Kristopher Kyle's institutional research funds. I am incredibly privileged to have been supported by these opportunities that made the project possible.

Finally, my wholehearted thanks go to my family: Masao (政夫), Hiromi (弘美), and Miku (美来). I could pursue my academic study in the U.S. because you believed in me. I am so lucky to be your son and brother.

This dissertation is dedicated to my parents and the memory of my grandmother, Yoko (洋子).

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Advanced writing proficiency is instrumental to academic success (Sparks et al., 2014). As such, research on academic writing has attracted much attention in English for Academic Purposes (EAP), which aims to reveal the nature of academic discourse (Hyland, 2005a) and has implications for pedagogy (Ferris & Hedgcock, 2013; Nesi & Gardner, 2012; J. M. Swales & Feak, 2000, 2012). There is a consensus that writing skills, as with other language skill areas, are a multifaceted, complex construct (Bachman, 1990; Bachman & Palmer, 1982; Canale & Swain, 1980; Ferris & Hedgcock, 2013; Hymes, 1972; Sparks et al., 2014), and teaching and assessment should take this complexity into account. In light of the multicomponent definitions of communicative competence (Bachman & Palmer, 1996; Canale & Swain, 1980), knowledge of lexis and grammar is important when translating ideas into words in the target language. This knowledge of lexis and grammar needs to be supported by knowledge of sociocultural norms around the context and the ways in which these features should be used to create *coherent* and *appropriate* discourse (sociolinguistic competence; Canale & Swain, 1980; see also Laughlin et al., 2015). A socio-cognitive approach to writing skills further highlights the process-oriented nature of writing by emphasizing the ability to plan, execute, and monitor non-linear writing processes (Deane, 2013; Kellogg, 1996). A genre-based approach to writing also emphasizes that the development and attainment of these skills should be discussed with the specific goals of the writing or situational context in mind (Burstein et al., 2016; Cotos, 2014). These perspectives suggest that defensible instruction and assessment of academic writing should encompass a range of approaches to conceptualizing writing skills while tailoring them to local contexts.

To materialize the multifaceted theoretical constructs of writing skills in local research and educational contexts at scale, one has to operationalize the constructs in observable measures (Bachman, 1990; Norris & Ortega, 2003).[1] Accordingly, a considerable amount of research on second language instruction and assessment has been devoted to identifying ways in which a given target construct manifests in language use, and the operationalization of such manifested behaviors in quantitative measures (e.g., Biber et al., 2020; Ellis & Barkhuizen, 2005; Housen et al., 2012; Norris & Ortega, 2003). To date, such collective efforts have resulted in a proliferation of frameworks to measure important linguistic constructs (often in an automatic manner), including lexical diversity (Jarvis, 2013a; Kyle, Crossley, et al., 2021) and sophistication (Kyle & Crossley, 2015; Laufer & Nation, 1995; Lu, 2012), syntactic complexity (Lu, 2011; Wolfe-Quintero et al., 1998) and sophistication (Kyle, 2016), and textual cohesion (Crossley et al., 2016; Graesser et al., 2004), to name but a few. These increasingly fine-grained operationalizations and automatic analyses of linguistic constructs have helped to achieve, for example, more precise measurement of complexity features under the Complexity, Accuracy, and Fluency (CAF) of language performance, often investigated in Task-Based Language Teaching (TBLT) and Instructed Second Language Acquisition (ISLA) research (see Michel, 2017; Spada, 2021).

While research on EAP, ISLA, TBLT, and language assessment has benefited from the ever-increasing ranges of operationalized measures of various constructs (and possibly because of such advances in this area), researchers are becoming increasingly aware that current

[1] In the present dissertation, I define the term *construct*, following the language assessment tradition, to mean "the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task" (Bachman & Palmer, 2010, p. 43). Thus, I do not specifically differentiate between cognitive (or psychological) constructs and performance-based constructs (see Chapelle, 2020 for a discussion of performance-type constructs).

measurement frameworks tend to focus excessively on the structural aspects of language production (e.g., frequency of lexical item, number of subordinate clauses) without much consideration of how these features are used in relation to the situational context and/or communicative goals of language performance (e.g., Eckes et al., 2016; Lu, 2021). To remedy this situation, more recent research has paid closer attention to the interplay between various complexification strategies and the contexts in which texts are produced (e.g., spoken vs written; text genres) (e.g., Biber et al., 2011; Polio & Yoon, 2018). Researchers on TBLT, although they have been using the CAF measurement framework for a long time (e.g., Norris & Ortega, 2009; Skehan, 2009), have increasingly explored degrees of communicative success in learners' performance more directly (e.g., Kuiken & Vedder, 2017; Revesz et al., 2014). More research now focuses on the interface between task-based performance and the pragmatic aspects of language use (e.g., Gilabert & Barón, 2018; Taguchi et al., 2021; Taguchi & Kim, 2018; Yasuda, 2015). Furthermore, research has also investigated how observed language features differ not only due to learners' proficiency but as a result of interaction with the nature of communicative demands of a task (e.g., Biber et al., 2014; Gray et al., 2019; Kyle & Crossley, 2016). To summarize, there seems to be a converging shift of research focus in EAP, ISLA, TBLT, and language assessment to closely examine the products of L2 writing (or speaking) and their differences across proficiency levels in relation to specific contextual parameters (e.g., communicative goals, genres, topics).

Despite this expanded research focus on the interactions of linguistic features and communicative contexts (e.g., Biber et al., 2014), the measurement frameworks for such research still remain predominantly structurally oriented. Unfortunately, very few frameworks to date have approached linguistic measurement by considering the potentially poly-functional nature of

linguistic features (see Lu, 2021). This is unfortunate because it does not allow us to draw conclusions about writers' ability to adapt to the communicative needs of tasks. An alternative fully functional approach would entail, for example, differentiating the specific functions in which adverbial clauses are used—e.g., temporal, conditional, or concessive (for an exception see Polio & Yoon, 2018). Arguably, such a functional measurement framework would provide essential information about language production, not only regarding the ways in which but also *the specific communicative purposes* for which writers use certain linguistic features in response to the communicative, functional demands of the immediate writing task (e.g., reporting research findings, making counterclaims, proclaiming their viewpoints).

Given the recent trend to expand the scope of second language measurement to include functional aspects of language use, the present dissertation explores one important functional construct often investigated in EAP research—namely, authorial evaluation, or evaluative language (e.g., Hunston & Thompson, 2000; Thompson & Alba-Juez, 2014; Xie, 2020). Thompson and Hunston (2000) define evaluation as a "broad cover term for the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about" (p. 5). It primarily concerns the interpersonal aspects of language use, which allows one to express opinions and maintain social relations with others (see also Biber & Finegan, 1988, 1989; Halliday & Matthiessen, 2014). Evaluation is pervasive in everyday language use (Hunston & Thompson, 2000; Martin & White, 2005), and particularly important in research on academic discourse (Hyland, 2005a) and media discourse (Martin & White, 2005). From an educational perspective, coverage of evaluative language use as a part of social and functional constructs of writing skills is of paramount importance for the design of

next-generation writing pedagogies and assessment (e.g., Burstein et al., 2016; Sparks et al., 2014).

Despite the significance of functional language use and interpersonal evaluative language in second language research, researchers have yet to agree on how this construct can be measured effectively (Carr, 2013; Lu, 2021). The present dissertation helps to address this gap in the literature—more precisely, it attempts to address the issues involved in measuring evaluative language in academic writing. This is done by combining existing frameworks and methodologies from discourse analysis, educational measurement and assessment, and computational linguistics/ natural language processing. The overarching goal of the dissertation is to demonstrate the development, evaluation, and application of an automated linguistic analysis tool that allows the identification and categorization of the rhetorical functions of evaluative language (see Chapter 2 for more precise definitions of terms and descriptions of the theoretical frameworks employed). As will be discussed in detail in Chapter 2, I draw on the theoretical framework of the engagement system from Appraisal Analysis (Martin & White, 2005), which is rooted in the Systemic Functional Linguistic (SFL) tradition of linguistics (Halliday & Matthiessen, 2014). Accordingly, the overall dissertation project consists of three independent studies, which are guided by the following research questions:

Study 1: What is the accuracy of an NLP pipeline in identifying engagement strategies (Martin & White, 2005) in academic English writing?

1. What is the baseline accuracy between trained human annotators?
2. How does the NLP pipeline perform compared to trained human annotators?

Study 2: How does the use of engagement strategies in university written assignments vary across:

1. disciplines,
2. genre families,
3. course levels,
4. assignment grades, and
5. the first language of the writer?

Study 3: What are the relationships between the use of engagement strategies and the assessed quality of timed L2 argumentative essays?

1. How are the rhetorical features of engagement distributed across two writing task types: email and argumentative essay tasks?
2. What is the relationship between the rhetorical features of engagement and essay qualities in argumentative and email writing tasks?
3. Do engagement strategies explain second language essay qualities above and beyond existing linguistic features at the levels of lexis, syntax, and cohesion?

The present dissertation is organized as follows. Chapter 2 reviews the current state of the literature, focusing on three theoretical frameworks often used in investigations of evaluative language in academic written English. This chapter then justifies the use of the engagement

system (Martin & White, 2005) to guide the present dissertation project. The chapter then briefly synthesizes empirical studies that have examined stance-taking across different academic writing contexts using the engagement framework. The chapter concludes with some methodological challenges faced in this discourse-analytic approach.

Chapter 3 reports on the development and validation of the Engagement Analyzer, focusing on (a) a corpus annotation project and (b) the underlying natural language processing (NLP) techniques used to train machine-learning models. The empirical part of this chapter includes the procedure for training models using the spaCy Python package (Honnibal et al., 2014/2020). Since this study (Chapter 3) is deemed instrumental to the overall dissertation project, considerable attention has been paid to making the report more transparent and comprehensive. For the annotation project, I draw on the stepwise annotation procedure from Fuoli (2018), which emphasizes the transparency, replicability, and reliability of an annotation project.

Using the developed NLP pipeline, Chapters 4 and 5 (Studies 2 and 3) showcase how the Engagement Analyzer, introduced in Chapter 3, can be used to describe the registers of university written assignments (Study 2, Chapter 4) and predict professional ratings of writing proficiency in timed L2 essays (Study 3, Chapter 5). Specifically, Chapter 4 explores the registers of university written assignments in terms of relative frequencies of engagement strategies. Chapter 5 reports on an empirical study investigating the relationships between engagement strategies and professional ratings of timed L2 writing. It demonstrates the benefits of using measures of engagement strategies alongside existing linguistic measures at the levels of lexis, grammar, and cohesion.

Although Chapters 3–5 can be considered independent in that they contain independent discussions of results, Chapter 6 summarizes the findings of all three studies and includes a statement of contributions for the present dissertation research. In the final section of Chapter 6, I reiterate some of the important limitations of the three empirical studies and outline several practical implications and future directions of research interfacing discourse analysis, second language acquisition, language assessment, and computational linguistics.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Introduction —Authorial Evaluation and Stance

Research suggests that academic writing includes a range of evaluative meanings and is far from objective (Hood, 2010; Hunston & Thompson, 2000; Hyland, 2000a). It has been shown that academic writing requires writers to negotiate epistemic knowledge claims with putative readers and use a range of resources to achieve these functions (Hood, 2010; Hunston & Thompson, 2000; Hyland, 2005b). Thus, research on authorial evaluation grew rapidly, and this led to the development of several theoretical frameworks often used in applied linguistics research (Biber, 2006a; Hyland, 2005a; Martin & White, 2005). As summarized by Xie (2020), evaluation concerns "writers' explicit or implicit encodings of their emotions of, viewpoints on, attitudes and positions towards entities or propositions in academic writing" (p. 1; see also Hunston & Thompson, 2000).

In research on authorial evaluation, researchers are often interested in two interrelated aspects of evaluative meaning—attitudinal and epistemic dimensions (Gray & Biber, 2012; Xie, 2020). The attitudinal dimension of evaluation attempts to capture the ways in which personal attitudes, emotions, and preferences on events and entities are encoded in language (Biber et al., 1999; Biber & Finegan, 1988, 1989; Gray & Biber, 2012). On the other hand, the epistemic dimension of evaluative meaning typically encompasses writers' levels of certainties, their commitment toward certain knowledge claims, and their acknowledgment of the sources of information (Gray & Biber, 2012; Martin & White, 2005; Xie, 2020).

For the reasons discussed in this chapter, the current dissertation project focuses on the epistemic dimension of evaluative meaning, mainly through the Appraisal framework in the systemic functional tradition (Martin & White, 2005). Based on this framework, I define stance as encompassing the epistemic dimension of how a writer positions themself concerning other voices and acknowledges recognition of alternative viewpoints (see Martin & White, 2005).

The overarching goals of the current chapter are twofold— (a) to provide an overview of three existing theoretical and methodological frameworks to investigate evaluative language in the English for Academic Purposes (EAP) literature, and (b) to review the state of the literature on evaluative meaning in the context of academic discourse and second language assessment. To this end, the chapter is organized as follows. In Section 2.2, I compare existing approaches to authorial evaluation. Specifically, I highlight three frameworks commonly used to investigate epistemic evaluation or stance. This section explains the motivation behind the selection of the engagement system (Martin & White, 2005) as the theoretical framework for the current dissertation project. I aim to show how the system of engagement relates to other approaches to the epistemic dimension of evaluative meaning, such as Biber's approach (e.g., Biber, 2006a) and Hyland's metadiscourse framework (Hyland, 2005a). In the following section (Section 2.3), I then introduce the system of engagement (Martin & White, 2005) and introduce some key terminology used throughout the current dissertation. In Section 2.4, I summarize existing research that has used the engagement system (Martin & White, 2005) to investigate evaluative language in academic discourse (Section 2.4). This is followed by a presentation of the benefits and drawbacks of using current qualitative discourse-analytic methods to study the engagement system (Section 2.6). The chapter then concludes with a summary outlining gaps in the literature that the present dissertation tries to fill.

2.2 Comparison of frameworks for stance-taking analysis

Traditionally, functional linguists have studied various linguistic phenomena which are relevant to authorial evaluation. These include but are not limited to modality (Palmer, 1986; Bybee and Fleischman, 1995), hedging (Brown and Levinson, 1987), evidentiality (Chafe, 1986; Chafe and Nichols, 1986), and affect (Ochs and Schieffelin, 1989). This research has tended to focus on specific aspects of evaluative language, and it has enabled researchers to propose more encompassing frameworks that attempt to capture a range of ways in which authors express their evaluation on topics through their writing (Hunston & Thompson, 2000; Gray & Biber, 2012). While such foundational research on specific aspects of evaluation (e.g., modality; Palmer, 1986) has undoubtedly contributed to the development of more encompassing approaches, given the applied focus of the current study, I limit the scope of the following review to the three most relevant frameworks in the context of English for Academic Purposes research. These are Biber's approach stance (Biber, 2006a; Biber & Finegan, 1988, 1989), Hyland's metadiscourse (Hyland, 2005a), and Martin & White's Appraisal (Martin & White, 2005).

2.2.1 Overview of the Comparison

Table 2.1 compares three approaches frequently used to investigate evaluative language in academic discourse studies (see Xie, 2020). These approaches share similar goals in their attempts to capture evaluative language use. However, as we shall see, they differ in their theoretical orientations, units of language they are interested in, and the scope and granularity of discourse categories they attempt to capture. The comparison also highlights that some frameworks are investigated predominantly through corpus linguistic tools, while others use qualitative discourse analysis.

Table 2.1

Comparison of three approaches to research on evaluative language.

Framework	Definition	Level	Functional/ Semantic category	Syntactic category	Typical methods
Stance (Biber et al., 1999)	Expressions of "personal feelings, attitudes, value judgments, or assessments" (Biber et al., 1999, p. 966)	Lexico-grammar	**A. Epistemic stance** 1) Marking certainty (or doubt), actuality, precision, or limitation adverbial 2) Marking the source or perspective of knowledge **B. Attitudinal stance** 3) Marking attitudes or evaluations 4) Marking personal feelings or emotions	- Stance adverbials - Stance (verb/ adjective/ noun) complement clauses - Modals and semi-modals - Stance noun + prepositional phrase - Premodifying stance adverb	Corpus-based analysis, typically through Biber tagger (Biber, 2006)
Metadiscourse (Hyland, 2005)	"cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community" (Hyland, 2005, p. 37)	Discourse semantics	**A. Interactive resources** 1) Transitions 2) Frame markers 3) Endophoric markers 4) Evidentials 5) Code glosses **B. Interactional resources** 1) Hedges 2) Boosters 3) Attitude markers 4) Engagement markers 5) Self-mentions	- "No simple linguistic criteria for identifying meta-discourse" (Hyland, 2005, p. 27)	1) Discourse analysis (Hyland, 2005) 2) Corpus-based pattern matching based on, e.g., Hyland (2005) (Yoon, 2017; Bax et al., 2019)

Table 2.1 (Cont'd)

Framework	Definition	Level	Functional/ Semantic category	Syntactic category	Typical methods
Appraisal (Martin & White, 2005)	**Attitudinal dimension (entity focused)** "how writers/ speakers approve and disapprove, enthuse and abhor, applaud and criticise, and with how they position their readers/ listeners to do likewise" **Epistemic dimension (proposition focused)** "how writers/ speakers construe for themselves particular authorial identities or personae, how they align or disalign themselves with actual or potential respondents, and how they construct for their texts an intended or ideal audience" (Martin & White, 2005, p. 1)	Discourse semantics	**A. Attitude** 1) affect 2) judgment 3) appreciation **B. Engagement** 4) monogloss 5) heteroglosses and their subcategories **C. Graduation** 6) force 7) focus	"a given attitude can be realised across a range of grammatical categories" (Martin & White, 2005, p. 10)	Discourse analysis (Martin & White, 2005)

2.2.3 Biber's stance analysis

Biber (1999) defines stance as expressions of "personal feelings, attitudes, value judgments, or assessments" (Biber et al., 1999, p. 966). Among the three frameworks reviewed, Biber's stance analysis (Biber et al., 1999; Biber, 2006a; Biber & Finegan, 1988, 1989) is the most linguistically oriented because it is primarily concerned with identifying linguistic realizations of evaluative meanings—how different lexico-grammatical devices are used for the expression of stance. In fact, Biber's approach is theoretically motivated by previous studies in functional linguistics and sociolinguistics, particularly the research on evidentiality (e.g., Chafe & Nichols, 1986), affect (e.g., Ochs & Schieffelin, 1989), and modality (e.g., Palmer, 2001).

Regarding a taxonomy of evaluative meaning, Biber et al. (1999) introduces a two-way distinction between attitudinal and epistemic stance (see Gray & Biber, 2012). These semantic categories of stance are explicitly linked to possible grammatical realizations in Biber et al. (1999). Attitudinal stance concerns personal feelings or the evaluation of an entity or event. For example, adverbial expressions such as *fortunately* or *interestingly* can be used to show evaluation and personal feelings. According to Biber (1999), attitudinal stance can be expressed through different lexico-grammatical devices, such as adverbials (e.g., *sadly*), Verb/Adjective/noun + complement clauses (e.g., *I* ***wish*** *I was there*), stance noun + preposition (e.g., *a* ***fear*** *of going out at night*), and modal verbs (e.g., *ought to*). On the other hand, epistemic stance pertains to a) marking of certainty, doubt, actuality, precision, or limitation or b) marking of the source or perspective of knowledge. The first category can be expressed through single-word or multiword adverbial expressions (e.g., *definitely, in fact, sort of*), verb/adjective/noun + complement clauses (e.g., *I am* ***certain*** *that we will do great),* stance nouns (e.g., *there is a* ***real***

possibility *of a split within the Lithuanian party*), and modal verbs (e.g., *might, may*). The second category can be expressed with adverbials (e.g., *according to*), verb + complement clauses (e.g., *the author* ***argued*** *that*), and nouns (e.g., a ***rumor*** of another oil-leak).

Regarding specific methods, Biber's (1999) stance analysis has been predominantly conducted using Biber tagger (Biber, 1984, 1988; Biber, Conrad, Reppen, et al., 2004) and subsequently developed open-source tools (Kyle, Choe, et al., 2021; Kyle et al., 2022; Nini, 2019). Biber tagger is a corpus tool developed by Doug Biber, which conducts a lexico-grammatical analysis of over 100 features (Biber, Conrad, Reppen, et al., 2004). It counts the occurrences of these features and returns their normalized frequency per 1,000 words. The features include a wide range of lexical and grammatical categories, such as pronouns, complement clauses, gerunds, modal verbs, semantic categories of nouns, adjectives, and verbs, etc. As such, stance features are often treated as a subset of semantic/ functional categories spanning different lexico-grammatical devices (Biber, 2006a). Table 2.2 provides some examples of stance expressions investigated in lexico-grammatical analyses (see Biber, 2006a).

As for the technical details, the approach typically uses combinations of surface lexical features and POS taggers (and syntactic parsers in some cases) followed by rule-based pattern-matching algorithms to identify lexico-grammatical features. In a recently reimplemented version of the tool (Kyle, Choe, et al., 2021; Kyle et al., 2022), the features are identified through combinations of POS, dependency tree representation, and surface features of items. In either implementation, raw input texts are tagged and parsed with the help of NLP tools, and identified patterns or lexical strings are looked up in a dictionary of candidate stance expressions (for a list see Biber, 2006a; Biber et al., 2004).

Table 2.2
Examples of stance expressions in Biber (2006).

Lexico-grammatical features	Epistemic		Attitude
	Certainty	*Likelihood*	
1. Stance adverbs	actually, certainly, in fact	apparently, perhaps, possibly	amazingly, importantly, surprisingly
2. Complement clauses			
2.1a Stance verb + that-clause	conclude, determine, know	believe, doubt, think	expect, hope, worry
2.1b Stance verb + to-clause		appear, happen, seem, tend, believe	need, want, intend
2.2a Stance adjective + that-clause	certain, clear, obvious	(un)likely, possible, probable	amazed, shocked
2.2b Stance adjective + to-clause	certain, likely, sure		happy, pleased, essential, important
2.3a Stance noun + that-clause	conclusion, fact, observation		hope, view
2.3b Stance noun + to-clause		tendency	failure

Note. Adapted from Biber (2006a).

2.2.4 Hyland's metadiscourse

Metadiscourse is arguably the most widely used framework when investigating non-propositional meanings in academic discourse (Ädel, 2006; Bouziri, 2021; Hyland, 2005a; Hyland & Jiang, 2022; for a review see Hyland, 2017). Essentially, metadiscourse concerns aspects of discourse that a writer or speaker weaves in addition to propositional meanings in order to help the recipient of the message understand what is being said (see Hyland, 2017). More formally, Hyland (2005a) defines metadiscourse as a "cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community" (p. 37). This definition capitalizes on the idea that written communication is a form of interaction

between writer and putative readers, and thus, the writer is expected to orient to the readers through various metadiscourse features. Accordingly, the metadiscourse includes not only the interpersonal but also textual dimension of discourse meanings, covering a wider range of constructs compared to the other two frameworks (Hyland, 2005a).

Metadiscourse is primarily a functional rather than a linguistic category (Hyland, 2005a). In other words, a metadiscoursal meaning can be expressed in an open set of lexico-grammatical expressions, and new items can be added to a shared repertoire in a discourse community. The function-oriented definition also means that metadiscourse can be expressed across different levels of linguistic construction such as lexis (e.g., *however*), phrasal constituents (e.g., *in fact*, *generally speaking*), clausal expressions (e.g., *it is possible that*), and whole sentences (Hyland, 2005a, 2017). For this reason, no grammatical categories are included in definitions of categories. Indeed, Hyland is clear that metadiscourse concerns open-class discourse categories, and there are "no simple linguistic criteria for identifying meta-discourse" (Hyland, 2005, p. 27). Interestingly, in discussing the non-propositional dimensions of discourse, Hyland (2005a) draws on the notion of meta-functions in Systemic Functional Linguistics (Halliday & Matthiessen, 2014). This fact arguably makes the metadiscourse framework relevant to (and compatible with) discourse analysis through the SFL framework.

Table 2.3 lists ten metadiscourse categories introduced by Hyland (2005a). In this framework, two main types of resources are identified. Interactive resources are those that help the reader navigate a text, including Transitions (e.g., *moreover*, *therefore*), Frame markers (e.g., *in this chapter*, *to conclude*), and Endophoric markers (e.g., *Fig. 4*, *in the next section*), among others. In this respect, it can be argued that interactive resources are closely related to textual

meta-functions in the SFL framework, which primarily concerns the function of a language to organize the discourse in terms of coherence and the orders of information (Halliday & Matthiessen, 2014 See Section 2.4.3 for the description of meta-functions), although no explicit links are made by Hyland (2005a). Interactional resources include resources to "involve the reader in the argument" (p. 49). These resources include both attitudinal and epistemic evaluative meanings (Hunston & Thompson, 2000). Epistemic dimensions are encoded in strategies where the writer displays their commitment to a proposition (i.e., booster) or mitigates a statement (i.e., hedges). Affective dimensions are covered through attitudinal markers. For this reason, interactional resources may be more closely related to the interpersonal meta-function in the SFL framework (See Section 2.2.3 for the description of the three metafunctions).

Regarding the methodological approach, the metadiscourse framework primarily employs discourse-analytic methods that allow the verification of functions in which a certain expression is used in discourse (Hyland, 2005a, 2017); however, some corpus-based tools have been developed in the context of language assessment and EAP research (i.e., Bax et al., 2019; Yoon, 2017a). In principle, context-specificity seems to be a key feature of metadiscourse, as Hyland (2005a, 2017) stresses it must be verified in context whether individual expressions function as metadiscourse. However, Hyland (2005a) also provides a list of candidate metadiscourse "expressions" as an Appendix. This appendix provided the basis for the corpus-based approach to metadiscourse in Text Inspector (Bax et al., 2019) and Authorial Voice Analyzer (Yoon, 2017a). For example, Text Inspector (Bax et al., 2019) uses over 300 examples of metadiscourse markers based on Hyland (2005a). The tool searches for occurrences of these 300 items in a user-input text and identifies any strings that match a given item in a predetermined list. Similarly, Yoon's (2017) Authorial Voice Analyzer allows the identification of interactive metadiscourse

markers in user-input texts. Focusing on interactive resources, it identifies 164 hedging expressions, 174 boosters, and 640 attitudinal markers, along with some expressions of self-mentions and reader pronouns, and directives (Yoon, 2017a). To summarize, although several corpus tools allow researchers to search for prototypical metadiscourse markers, full-blown metadiscourse research still requires a discourse-analytic method (Hyland, 2017).

Table 2.3
Hyland's (2005) Metadiscourse framework.

Category	Function	Examples
Interactive resources	Help guide the reader through a text	
Transitions	Express semantic relation between clauses	in addition/ thus/ and/ furthermore/ in contrast
Frame markers	Refer to discourse acts, sequences, or text stages	to conclude / my purpose is / to begin with
Endophoric markers	Refer to information in other parts of the text	noted above / in section 2 / see Table
Evidentials	Refer to sources of information from other texts	according to X / Z argues / (Y, 1999)
Code glosses	Help readers grasp the meanings of ideational material	namely / in other words / e.g., / such as
Interactional resources	Involve the reader in the argument	
Hedges	Withhold a writer's full commitment to a proposition	might / possibly / perhaps
Boosters	Emphasize force or writer's certainty in a proposition	in fact / definitely / it is clear that
Attitude markers	Express writer's attitude to a proposition	unfortunately / I agree / surprisingly
Engagement markers	Explicitly refer to or build a relationship with the reader	consider X / note that /
Self-mentions	Explicit references to author(s)	I / we / my / our

Note. Adopted from Hyland (2005a, p. 49, Table 3.1).

2.2.3 Martin & White's Appraisal Framework—A Systemic Functional Approach

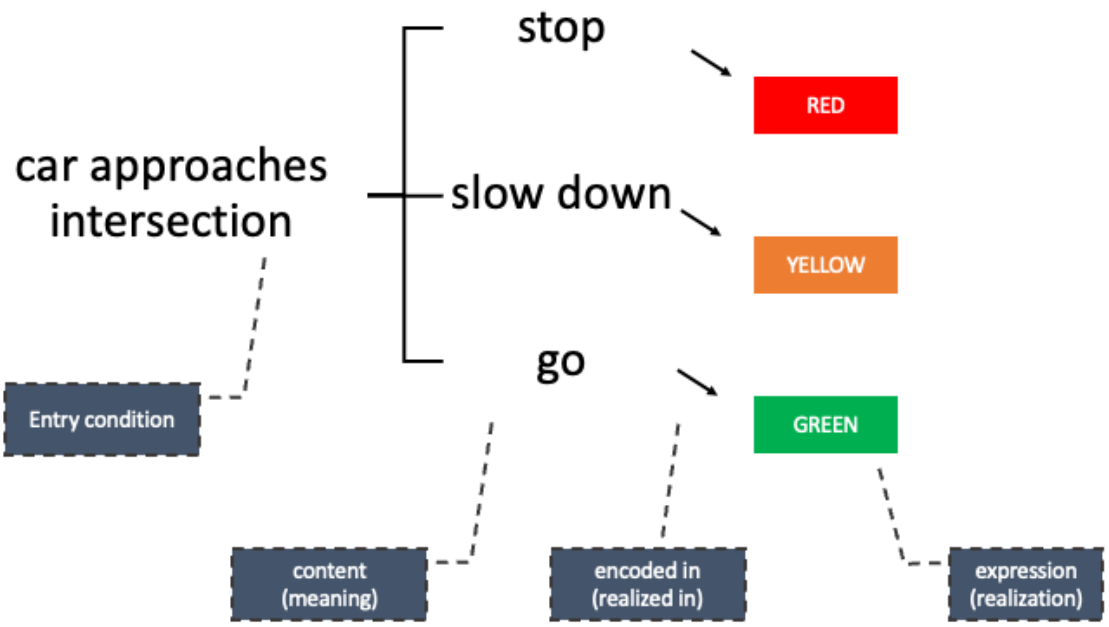
The third framework covered in this review is Martin and White's Appraisal framework (Martin & White, 2005). It is an all-encompassing functional framework of interpersonal meanings developed primarily by Martin and White (Martin & White, 2005; White, 2003). The appraisal framework comprises three dimensions: Attitude, Engagement (the focus of the current dissertation), and Graduation. It concerns aspects of language for "the negotiation of social relationships" (Derewianka, 2007, p. 143). It concerns the discourse semantics of texts rather than the grammatical categories to achieve them. It is important to note that the engagement system in Martin and White (2005) has nothing to do with Hyland's (2005a, 2005b) engagement marker.

Before discussing the framework in more detail, a few notes on the theoretical underpinnings of Systemic Functional Linguistics are appropriate. A central tenet of SFL is that it views language as a semiotic resource or a tool for meaning-making (Eggins, 2004; Halliday & Matthiessen, 2014). Specifically, SFL posits that language affords the means to achieve three metafunctions that are simultaneously at play in any human communication: namely, ideational, interpersonal, and textual metafunctions. First, the ideational metafunction concerns the meaning of language that construes human experience and represents the world around us. Language allows one to describe entities and events to another individual who has not perceived or experienced them first-hand. Second, the interpersonal metafunction of language allows one to establish and maintain social relationships with others. It states that language enacts social relationships, in that it provides means for asking questions, giving orders, making offers, and expressing our evaluations, attitudes, and appraisals. Third, the textual metafunction concerns the

functions of language that enable and facilitate communication. Essentially, it is the function of language to organize discourse into coherent pieces so that the recipient of a message can comprehend it. To summarize, SFL considers that the features of (academic) language enable researchers to exchange novel ideas or findings (i.e., ideational) in a convincing manner and, importantly, without upsetting the audience (i.e., interpersonal) or making the reader feel lost (i.e., textual).

Another important tenet of SFL is that it attempts to model functional choices using a series of paradigmatic taxonomic networks of related semiotic resources, also known as system networks (Halliday & Matthiessen, 2014). Figure 2.1 (adapted from Eggins, 2004) illustrates the concept of this system network using a traffic signal as an example. Typically, the semiotic nature of traffic signals becomes relevant when a car approaches an intersection where a signal regulates the traffic (this is expressed as an entry condition). The function/ meaning of a traffic signal can be expressed in a paradigmatic network of three values: namely, stop, slow-down, or go. The *paradigmatic* relations among these functions are important—three discrete functions make up the meaning potential of a traffic signal, and the branches indicate that only one choice from the system network is possible at a time. Also important is the fact that this is a system of semiotic resources, showing how each choice of meaning is encoded in (or realized in) a concrete expression. In summary, the SFL approach uses these system networks to theorize and model the meaning potentials of language in each domain of ideational, interpersonal, and textual metafunctions (Eggins, 2004; Halliday & Matthiessen, 2014).

Figure 2.1
The semiotic system of traffic lights



Note. Adapted from Eggins (2004, p. 14).

With the two central tenets of the SFL approach in mind, Figure 2.2 provides the overarching system network of appraisal resources (Martin & White, 2005). It shows that the appraisal framework encompasses three domains of interpersonal meanings—Engagement, Attitude, and Graduation. These three domains are consistent with existing frameworks of evaluative language in that Attitude and Engagement correspond to affective and epistemic dimensions, respectively (Biber et al., 1999; Biber & Finegan, 1988, 1989; Hunston & Thompson, 2000), and Graduation essentially relates to a process of intensification (e.g., Labov, 1984).

Concerning the affective dimension, the ATTITUDE subsystem attempts to describe a system network of "shared feelings and values" and the "sharing of emotions, tastes and

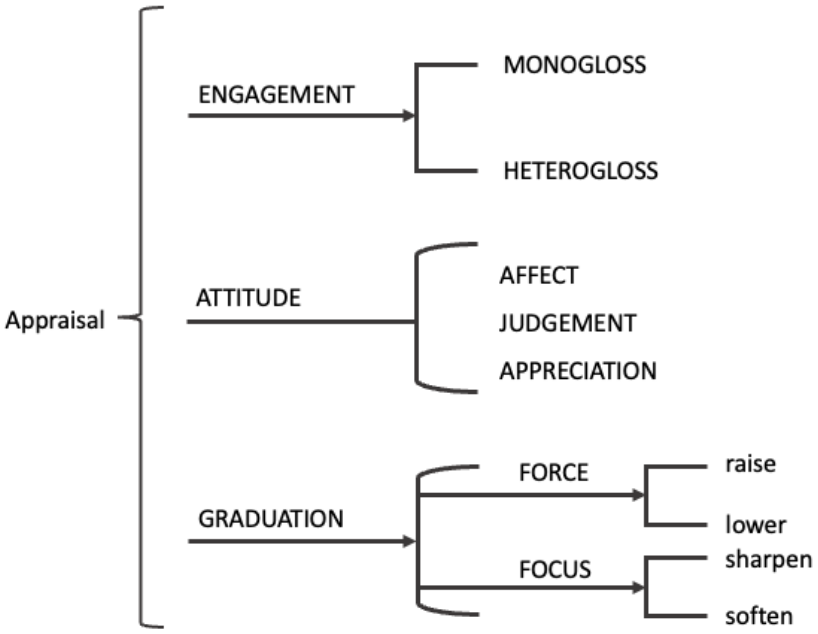
normative assessments" (Martin & White, 2005, p. 1). There are three "semantic regions" (Martin & White, 2005, p. 42) proposed that roughly relate to emotion, ethics, and aesthetics. Affect covers expressions of emotions or feelings, including but not limited to (un)happiness, (in)security, (dis)satisfaction, and sadness (Martin & White, 2005). Judgment concerns the expressions used to admire, criticize, and praise certain behaviors. It includes semantics relating to social esteem (e.g., specialty, capability, dependability) and social sanction (e.g., honesty, morality). For example, judgment can be realized with the following example adjectives: *wrong, cruel*, *kind*, *humane*, *skillful*, etc. (Martin & White, 2005). Appreciation includes the semantic field of evaluating someone's work or natural phenomena. For example, this includes the semantic process where a speaker or writer describes something as *dull*, *dramatic*, *beautiful*, *consistent*, *flawed*, *elegant*, *plain*, *profound,* or *shallow*.

ENGAGEMENT is concerned with the epistemic dimension of evaluation, or constructions of "authorial identities or personae" (Martin & White, 2005, p. 1), focusing on "how they align or disalign themselves with actual or potential respondents, and […] how they construct for their texts an intended or ideal audience" (Martin & White, 2005, p. 1). This system of engagement is inspired by Bakhtin's dialogism, where all statements in some way invoke alternative positions that have been observed previously in the speech community and, in this sense, are considered potentially "dialogic" in nature (Bakhtin, 1981; as cited in Martin & White, 2005). From this perspective, the system of engagement posits that a speaker or writer has discourse semantic choices when formulating their epistemic claims—by showing their recognition of potential alternative viewpoints on a matter of discussion (heterogloss) or by disregarding alternatives (monogloss). Details of the engagement system are covered in section 2.3 as it is the focal framework for the current dissertation study.

GRADUATION concerns how speakers and writers "up-scale" or "down-scale" the discourse semantic values of Attitude and Engagement dimensions (Martin & White, 2005). The appraisal framework posits two Graduation strategies that concern the scalability of Attitude or Engagement: FORCE and FOCUS. FORCE concerns grading with respect to intensity or amount (Martin & White, 2005). This process often involves scaling a process or quality (e.g., *slightly*, *extremely*, *somewhat*). It can also involve quantifications of numbers (e.g., *few*, *many*), mass (e.g., *huge*, *small*), proximity (e.g., *recent*, *distant*), and distribution (e.g., *long-lasting*, *widespread*). In contrast, FOCUS deals with grading based on prototypicality and preciseness. For example, a writer can soften the locution using expressions often recognized as "hedges", including "sort of" or "kind of" (e.g., *they* ***sort of*** *play jazz*; Martin & White, 2005). Alternatively, a writer can enhance prototypicality by adding attributive adjectives (e.g., *they play* ***real*** *jazz*).

To summarize, the appraisal framework provides a fully functional system network of interpersonal meanings of language—how writers express their feelings and attitudes (Attitude), how they position themselves among others (Engagement), and how they grade their attitudinal and epistemic evaluations (Graduation).

Figure 2.2
Overview of the Appraisal System



Note. Adapted from Martin and White (2005); There are two kinds of system architecture proposed in this network. The first is bracket-like representations (as in ENGAGEMENT, ATTITUDE, and GRADUATION under Appraisal and AFFECT, JUDGMENT, and APPRECIATION under ATTITUDE). This means that choices can be independent and happen simultaneously in discourse. For example, a given linguistic construction can signal a value from Engagement and Graduation simultaneously (e.g., *there is* ***mounting evidence*** [ATTRIBUTE and FORCE] *that*). On the other hand, tree-like representations denote paradigmatic choices (as in distinctions between MONOGLOSS or HETEROGLOSS in the ENGAGEMENT branch and sharpen or soften in FOCUS). This implies that when an utterance has a heteroglossic value (i.e., recognizes alternative viewpoints), the same utterance cannot be interpreted as monoglossic (i.e., disregards alternative viewpoints).

Regarding the methodological approach, the appraisal framework typically employs a discourse-analytic approach (Martin & White, 2005; see Xie, 2020). In-depth discourse analysis allowed researchers to focus on the meaning in discourse more directly, rather than focusing on individual linguistic realizations (e.g., Hood, 2010; Lancaster, 2014; Xu & Nesi, 2019). In particular, qualitative textual analysis allows analysts to uncover incongruent types of

grammatical realizations (see Hood, 2010). This includes rank-shifted expressions (Halliday & Matthiessen, 2014), such as grammatical metaphor (e.g., *he is* ***excellent*** [adjective; congruent] → *his* ***excellence*** [noun; incongruent, rank-shifted]). Discourse analysis also allows finer distinctions between internal (text-internal) versus external (real-world) conjunctive relations (Martin et al., 1997; Martin & Rose, 2007). In this sense, the underlying functional orientation of the framework is shared with the metadiscourse framework (Hyland, 2005a), where not only individual items but also co-textual information (how an item is used in discourse) are critical in determining whether they are used as appraisal resources.

2.2.3 Comparison Summary

The comparison presented above has revealed a number of (dis)similarities among the three frameworks. All three include two separate dimensions of evaluative meanings—epistemic and attitudinal—in their taxonomies. Biber's stance and Martin and White's (2005) appraisal framework focus on evaluative meanings, while Hyland's metadiscourse includes textual dimensions of discourse (i.e., interactive resources) in addition to evaluative aspects (i.e., interactional resource). Methodologically, Biber's approach relies on corpus searches, whereas the appraisal framework is mainly discourse-analytic, attending to specific semantic values in discourse more than grammatical realizations. Hyland (2005) proposed metadiscourse as a discourse-analytic framework, while there is a recent trend to move toward corpus-based analysis using published lists of metadiscourse markers (Bax et al., 2019; Yoon, 2017a).

Beyond these, there are a few crucial differences that motivated the selection of the engagement system in appraisal as the focal framework for this study. First, Hyland treats

evidentials as an interactive resource (cf. textual features), while this is part of the epistemic dimensions of evaluative meaning in the other two frameworks and in other widely accepted models of evaluative language (e.g., Gray & Biber, 2012; Hunston & Thompson, 2000). Compared to the Appraisal framework (Martin & White, 2005), Biber's framework attends more to the decontextualized lexical semantics of individual items (e.g., certainty, likelihood) and does not speak to discourse categories that are often critical in the analysis of evaluative language and its application to EAP (e.g., self-sourced or other-sourced argument; see Charles, 2006). This second point is essential when analyzing stance in academic discourse because writers are expected to negotiate epistemic claims with respect to possible alternative viewpoints on a given matter (e.g., Hood, 2010; Hyland, 2005a). The same is true for the metadiscourse framework, where the basic distinction between averral and attribution (Charles, 2006; Hunston & Thompson, 2000) is not encoded in a straightforward manner in the framework. This may limit discourse analysis to being somewhat vague in terms of a writer's relative stance on a specific topic.

To illustrate these (dis)similarities, Table 2.4 presents illustrative textual analyses I conducted following the three frameworks. For this illustration, I used an excerpt from a published paper (Derewianka, 2007) to showcase the differences in the three frameworks. As can be seen, while the three frameworks target similar aspects of discourse, they differ significantly in terms of the granularities of the analyses. In particular, discourse-oriented approaches—e.g., Martin and White (2005)—make finer-grained distinctions regarding discourse semantic values (e.g., negative versus positive), in addition to whether or not an expression is categorized as an attitudinal expression (Biber, 2006a). Such finer-grained information is arguably relevant in

analyses of evaluative language. For this reason, I will focus on the engagement system from the appraisal framework in the current dissertation project.

2.3 The engagement system

Figure 2.3 presents the engagement system in the appraisal framework from Martin and White (2005). As mentioned earlier, the appraisal framework is conceptualized within Systemic Functional Linguistics (SFL) and thus provides a systemic network of discourse semantic choices pertaining to evaluation. In this framework, the engagement system attempts to provide a paradigmatic taxonomy of discourse choices available to writers when they make knowledge claims in their writing. From an analytical perspective, subordinate categories are seen as more granular discourse strategies inheriting the core values of immediate parents (e.g., ENTERTAIN and ATTRIBUTE are considered sister categories because they are both expanding strategies). Throughout the current dissertation, the term engagement system refers to the entire systemic taxonomy (as described in Fig. 2.3) that describes the paradigmatic functional choices available to writers. The term engagement strategy refers to a discrete discourse strategy such as ATTRIBUTE, PRONOUNCE, or COUNTER. Finally, I use engagement resources to refer to the lexico-grammatical means which writers employ to realize engagement strategies in the immediate context.

The current study focuses on eight main strategies of engagement from Martin and White (2005), which are color-coded in Figure 2.3, plus JUSTIFY, which is included in an earlier version (White, 2003) as well as in some empirical studies (Lam & Crosthwaite, 2018). The details of the engagement system are described below, focusing on the distinction between monogloss and heterogloss, expansion and contraction strategies, and the discrete engagement strategies I focus on in this study.

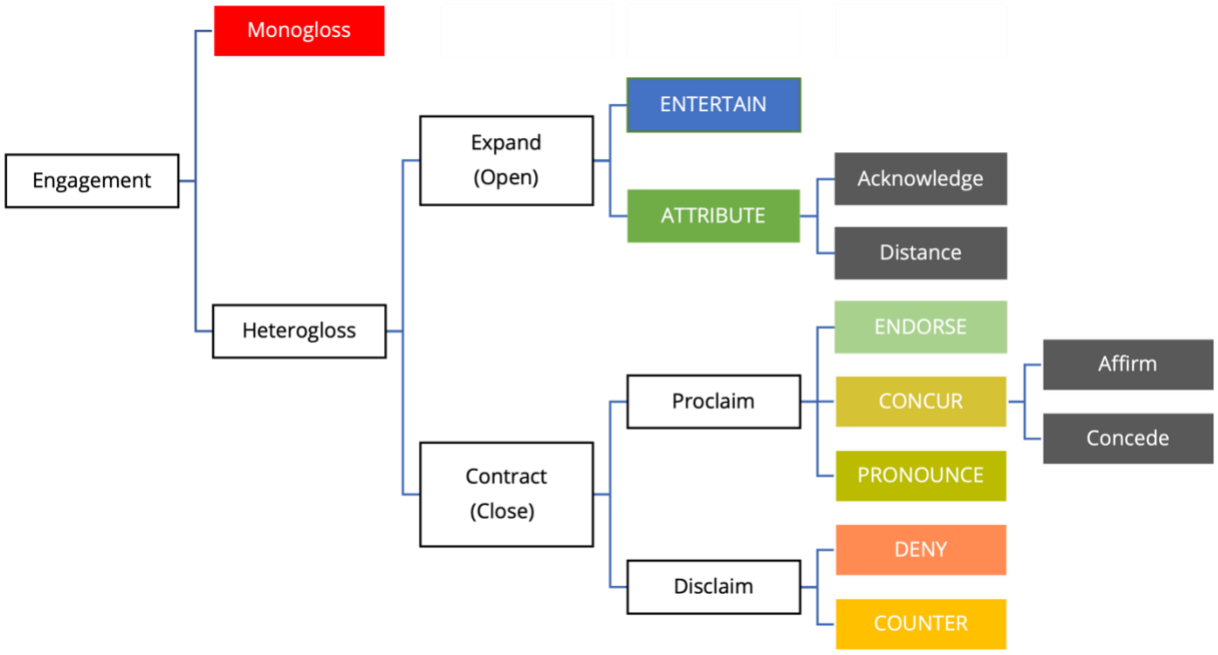
Table 2.4

Illustrative Textual Analysis with Three Frameworks.

Biber's Stance analysis (2006a)	Hyland's Metadiscourse (2005a)	Martin and White's Appraisal (2005)
Breitman, in his article Plans for the Final Solution, refers to the controversy about the origins of the Holocaust as the "intentionalist-functionalist" debate; that is one between those who **think [likelihood verb]** the Holocaust was a preconceived Nazi plan, and those who **think [likelihood verb]** it was improvised hastily [Attitudinal Stance], notably [Attitudinal Stance] after early German victories in the Soviet Union in mid-1941.	Breitman, **in his article Plans for the Final Solution [Evidentials]**, refers to the controversy about the origins of the Holocaust as the "intentionalist-functionalist" debate; **that is [Code gloss]** one between those who think the Holocaust was a preconceived Nazi plan, and those who think it was improvised hastily [Attitudinal Marker], notably [Attitudinal Marker] after early German victories in the Soviet Union in mid-1941.	Breitman, **in his article Plans for the Final Solution [ATTRIBUTE]**, **refers to [ATTRIBUTE]** the controversy about the origins of the Holocaust as the "intentionalist-functionalist" debate; that is one between those who **think [ATTRIBUTE or MONOGLOSS[*1]]** the Holocaust was a preconceived Nazi plan, and those who **think [ATTRIBUTE or MONOGLOSS[*1]]** it was improvised hastily [Appreciation: Negative], notably [Appreciation: Positive] after early German victories in the Soviet Union in mid-1941.
Breitman, in contrast to Browning, is very much an 'intentionalist', **arguing that [Stance verb + that-clause]** the 'murderous intentions' of Hitler, Himmler and other key Nazis were well underway before the invasion of the Soviet Union.	Breitman, **in contrast to [Transition]** Browning, is **very much [Booster]** an 'intentionalist', arguing that the 'murderous intentions' of Hitler, Himmler and other key Nazis were **well [Booster]** underway before the invasion of the Soviet Union.	Breitman, in contrast to Browning, **is [MONOGLOSS]** *very much [Force: up-scale]* an 'intentionalist', **arguing [ATTRIBUTE]** that the 'murderous intentions' of Hitler, Himmler and other key Nazis were *well [Focus]* underway before the invasion of the Soviet Union.
Whilst he **admits that [Stance verb + that-clause]** many of the Nazi documents on this issue are "inexact", he **suggests that [Stance verb + that-clause]** this is not because the Holocaust plan itself was **uncertain [certainty adjective]**, but rather [Attitudinal Stance] because the Nazi leadership **wanted [Desire verb]** to "conceal" and "veil" its real intentions from others (p. 271).	**Whilst [Transition]** he admits that many of the Nazi documents on this issue are "inexact", he **suggests [hedge]** that this is not because the Holocaust plan itself was uncertain, **but [Transition]** rather because the Nazi leadership wanted to "conceal" and "veil" its **real [Booster]** intentions from others **(p. 271) [Evidentials]**.	Whilst he **admits [ATTRIBUTE]** that *many [Force: up-scale]* of the Nazi documents on this issue are "inexact" [Appreciation: Negative], he **suggests [ATTRIBUTE]** that this is **not [DENY]** because the Holocaust plan itself was uncertain [Appreciation: Negative], **but rather [COUNTER]** because the Nazi leadership wanted to "conceal" and "veil" its *real [Focus: sharpen]* intentions from others (p. 271).

Note. These textual analyses were conducted manually by the author; the excerpt is taken from Derewianka (2007, pp. 156–7), which was introduced as an example of early tertiary writing. [*1] Regarding the functional categories of *think* in these contexts under the Appraisal framework (Martin & White, 2005), I suggest two possible interpretations—ATTRIBUTE or MONOGLOSS. ATTRIBUTE interpretation is possible given that the writer makes explicit reference to two positions on the "intentionalist-functionalist" debate under discussion. In other words, the first proposition, "the Holocaust was a preconceived Nazi plan," is not that of the author but one of the positions in the debate. According to this reading, it is possible to argue that the mental verb *think* has been used to ATTRIBUTE each proponent of the positions in the debate. At the same time, the analysis is complicated by the fact that these positions are introduced as presupposed contents. In other words, while the writer does make reference to (and report) the two external positions, this attribution is nested in the author's averral (i.e., the author's self-sourced argument concerning the two positions). This multilayered complexity of evaluative language has been repeatedly pointed out in the literature (e.g., Hunston, 2000, p. 179). Such complexities in the annotation of discourse features of engagement will be made clearer in Study 1 (e.g., Fuoli, 2018).

Figure 2.3.
The engagement system (Adapted from Martin & White, 2005).



2.3.1 Monogloss or Heterogloss

In the system of engagement, the top-level distinction can be made between monogloss and heterogloss. This binary decision concerns whether the immediate utterance recognizes alternative positions. In monogloss (mono = 'single', gloss = "commenting on a text," Etymonline, n.d.), the writer does not recognize any alternative views and presents the idea/ event as if it is a fact (e.g., "The banks have been greedy"; Martin & White, 2005). On the other hand, a heteroglossic utterance includes various ways in which writers display their recognition of possible alternatives to the matter under discussion (e.g., "I *speculate* that the banks have been greedy", "I *read somewhere* that the banks have been greedy", "It is *unlikely* that the banks have been greedy", etc.). Although recent studies have proposed subcategories under monogloss, such as factual statement, assertion, and presupposition (S. H. Lee, 2017), to the best of my knowledge, these subcategories have not yet been fully incorporated into the system of

engagement. In this study, I will not make any further distinctions between sub-monoglossic categories for the sake of simplicity (but the benefits of subcategories will become evident in later stages in the dissertation; see Chapter 3).

2.3.2 Expansion or Contraction

A heteroglossic statement can be distinguished in terms of whether the writer makes some allowances for alternative standpoints in the immediate discourse (expansion) or does not allow such room for negotiation (contraction). Expansion strategies are those that open up dialogic spaces for alternative viewpoints. This can be done using strategies such as ENTERTAIN or ATTRIBUTE (see Section 2.3.3). On the other hand, writers can close down dialogic spaces by disclaiming any alternative claims or proclaiming their viewpoints (see Section 2.3.4).

2.3.3 Dialogic expansion—ENTERTAIN or ATTRIBUTE

Expansion strategies include discourse strategies that (a) increase the tentativeness to the statement (ENTERTAIN) and that (b) attribute the idea to external sources (ATTRIBUTE). In ENTERTAIN, a writer can use lexico-grammatical items such as modal verbs (*can*, *may*) and mental verbs (I *believe*) to entertain other possible alternatives. ENTERTAIN is closely related to but still distinct from hedges in the metadiscourse literature (Hyland, 2005a), and low-certainty expressions (e.g., *perhaps*) and subsets of communication verbs (e.g., I *think*) in Biber's corpus-based approach (Biber, 2006a). In ATTRIBUTE, writers mention what is reported in (un)identified external sources (e.g., the paper *mentioned*, *it is believed* that). This is related to evidentials in the metadiscourse literature (Hyland, 2005a), and subsets of communication verbs

and adverbial expressions in Biber's approach (Biber, 2006a). In Martin and White's original (2005) system, ATTRIBUTE can be broken down further in terms of Acknowledge or Distance, depending on whether the writer takes a neutral stance on the reported content (Acknowledge) or show some skepticism toward it (Distance). However, given that previous research often does not differentiate these choices (e.g., Lancaster, 2014; Wu, 2007), I use ATTRIBUTE as a category that subsumes both Acknowledge and Distance instances.

2.3.4 Dialogic contraction—Disclaim or Proclaim

In contrast to expansion strategies, contraction include discourse strategies where writers attempt to close the dialogic space for discussion. This is done by either rejecting other viewpoints (Disclaim) or bolstering their own views (Proclaim). In disclaim, writers can DENY the reliability of a particular point of view (e.g., That is *NOT* correct) or COUNTER the alternative ideas (e.g., *Although the paper may be right*, there is another possibility). In proclaim, writers attempt to enhance the validity of their views by (a) formulating the locution assuming that it will be easily accepted by putative readers (CONCUR; e.g., *as you know*, *of course*, *surely*), (b) showing extra commitment to the validity of their views (PRONOUNCE; we *conclude* that), or (c) presenting other's perspective/ data/ claims as correct and reliable and aligning to them (ENDORSE).

2.3.5 Auxiliary strategy—JUSTIFY

Although not included in Martin and White's (2005) system, JUSTIFY is another strategy considered in White's (2003) system. According to White (2003), JUSTIFY concerns "formulations which construe a particular type of consequentiality, namely those by which non-

'factual' propositions (for example, attitudinal evaluations, directives/ recommendation, predictions and so on) are justified, substantiated or otherwise argued for" (p. 274). Typical JUSTIFY resources include conjunctions such as *therefore*, *thus*, *accordingly*, *because*, and *for this reason* (White, 2003, p. 274).

2.3.6 Summary of the engagement system

Table 2.5 lists eight plus one discrete engagement strategies considered in the current dissertation. The overall network of Martin and White (2005) remains intact, the only minor modifications are the addition of JUSTIFY strategy from White (2003) and the use of ATTRIBUTE and CONCUR as categories subsuming finer-grained distinctions. As we will see below, this nine-category classification is compatible with previous empirical studies that examined engagement strategies and their resources in academic writing (Chang & Schleppegrell, 2011; Lancaster, 2014; Wu, 2007). In the next section, I will review some empirical studies that used the engagement system to investigate evaluative language in L1 and L2 writing.

Table 2.5
Summary of engagement strategies (adapted from Martin & White, 2005; White, 2003).

Macro strategy	Engagement strategy	Description	Examples of prototypical lexico-grammatical realizations (see Chang & Schleppegrell, 2011)
Contraction	Disclaim: Deny	An utterance which invokes a contrary position but which at the same time rejects it directly. The contrary position is hence given very little dialogic space.	• Negative particles (e.g., *not, never*)
Contraction	Disclaim: Counter	An utterance which expresses the present proposition as replacing and thus 'countering' another proposition which would have been expected.	• Conjunctions (e.g., *but*) • Adverbials (e.g., *however*) • Adverbial clauses (e.g., *although*)
Contraction	Proclaim: Concur	An utterance which shows a writer's expectation/ assumption that putative readers will agree with the proposition and/or have the same knowledge.	• Adverbials (e.g., *indeed, of course*) • Adverbial clauses (e.g., *as one expects*) • Display questions; tag questions
Contraction	Proclaim: Pronounce	An utterance which expresses a strong level of writer commitment through the author's explicit emphasis and interpolation, thereby closing down the dialogic space.	• Mental/Communication verbs (e.g., I *contend*, we *conclude*, I *propose*) • Emphatic do (e.g., I *do* believe that) • Modal attributes (e.g., *it is evident* that)
Contraction	Proclaim: Endorse	An utterance which refers to external sources as warrantable, undeniable, and/or reliable. It expresses the writer's alignment with and endorsement of an attributed proposition. As such, the dialogic space is somewhat narrowed.	• Reporting verb (e.g., Kyle (2020) *demonstrated* that)

Table 2.5 (Cont'd)

Macro strategy	Engagement strategy	Description	Examples of prototypical lexico-grammatical realizations (see Chang & Schleppegrell, 2011)
Expansion	Entertain	An utterance which indicates the author's position but as only one possibility amongst others, thereby opening up dialogic space.	• Modal verbs (mainly epistemic and deontic modals; e.g., *may*, *would*; Palmer, 2001) • Mental/Communication verbs (e.g., I *think*/*suppose*, we *suggest*) • Adverbials (e.g., *perhaps*, *probably*) • Adverbial clauses (e.g., *unless*, *when*, *if*) • Modal attributes (e.g., *it is likely that*) • Evidentials (e.g., *seem*, *apparently*)
Expansion	Attribute	An utterance which signifies dialogic space as the writer attributes the proposition to an external source.	• Adverbials (e.g., *reportedly*) • Reporting verbs (e.g., *they argue*/ *believe*)
Monogloss	Monogloss	An utterance which does not employ any value of engagement. Such an utterance ignores the dialogic potential in an utterance.	• Present-tense verbs • Lacks of any other engagement strategies
Auxiliary	Justify	An utterance which engages in persuasion through justification or substantiation.	• Adverbials (e.g., *therefore*, *for this reason*) • Conjunctions (e.g., *because*)

2.4 Engagement in academic writing

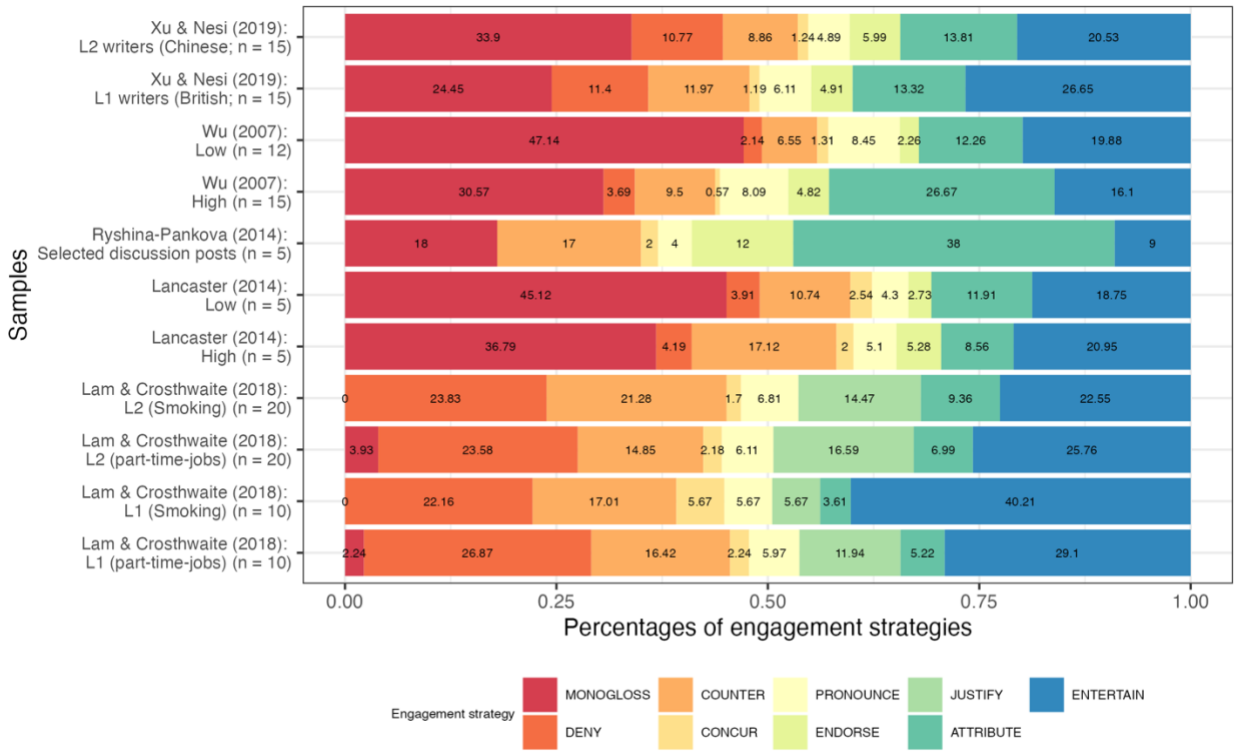
The system of engagement has been used to reveal patterns of stance-taking in peer-reviewed journal articles (Chang & Schleppegrell, 2011; Hood, 2010; Xu & Nesi, 2019), university written assignments (Lancaster, 2014; Wu, 2007), and online blog posts completed as part of coursework (Ryshina-Pankova, 2014), among others. This research can be categorized into two main categories regarding their goals and methodologies. The first category focuses primarily on the qualitative descriptions of how engagement strategies are realized in academic discourse, enabling a more in-depth understanding of the engagement system and the resources used to realize the system in academic discourse in general (Chang & Schleppegrell, 2011; Hood, 2010). The second type of study uses qualitative analysis followed by a quantitative examination to uncover how groups differ in terms of their relative proportions of engagement strategies (e.g., Lancaster, 2014; Wu, 2007). In this section, I mainly focus on the latter approach and summarize the overall patterns and a few diverging findings across reported studies in this domain.

By way of a brief synthesis, Figure 2.4 plots the relative proportions of discrete engagement strategies reported in previous studies, calculated from either raw counts or normed frequencies. In this line of research, researchers are interested in examining how groups of writers differ in their uses of engagement strategies (Lam & Crosthwaite, 2018; Lancaster, 2014; Wu, 2007; Xu & Nesi, 2019). To date, research has examined published applied linguistics research articles by L1 and L2 writers (Xu & Nesi, 2019), essays written by university students in a Geography class in Singapore (Wu, 2007), blog posts as part of university course assignments (Ryshina-Pankova, 2014), policy analysis writing submitted for a 400-level Economy class at a U.S. university (Lancaster, 2014), and timed argumentative essays (Lam &

Crosthwaite, 2018) sampled from a large-scale corpus, the International Corpus Network of Asian Learners of English (Ishikawa, 2013).

Figure 2.4

Relative proportions of engagement strategies as reported in previous studies.



As can be seen from Figure 2.4, the relative proportions of engagement strategies appear markedly different across samples, although some strategies are more frequent than others overall. For example, expansion strategies (ENTERTAIN and ATTRIBUTE) tend to occupy more than a quarter of the entire distribution in each study, approaching half in Wu (2007), Ryshina-Pankova (2014), and Lam and Crosthwaite (2018). Also, MONOGLOSS tends to occur 18% of the time or more, except in the Lam and Crosthwaite (2018) study. On the other hand, writers across these studies tend to use fewer Proclaim strategies—CONCUR, PRONOUNCE,

and ENDORSE. In particular, CONCUR occurs in only 0.5-5.67%, suggesting that this strategy is scarce across different situational contexts examined thus far.

More direct comparisons can be made within individual studies investigating differences across sub-populations (e.g., Lancaster, 2014; Wu, 2007). Lancaster (2014) sampled five lower-graded and five higher-graded essays from a data set of 194 authentic policy analysis writing samples from a 400-level Economics course. These essays were sampled to exclude essays scored around the average range. Similarly, Wu (2007) selected essays evaluated as A or C letter grades in an introductory Geography course at a Singaporean university. In these focused analyses of extreme cases, they found less frequent use of a MONOGLOSS strategy and more frequent use of COUNTER in higher-graded essays. There was a similar trend in Xu and Nesi's (2019) study, where MONOGLOSS was less frequent in L1 writers' published manuscripts than L2 authors' ones. There seems to be a tendency for more frequent expansion strategies in higher-graded essays in Wu (2007) and L1 authors' published manuscripts in Xu and Nesi (2019). These patterns suggest that the relative distributions of discrete engagement strategies may vary depending on the genre of writing, the assessed quality of writing, and writers' L1 backgrounds. This synthesis also suggests that the systematicity of such patterns may require further testing before more decisive conclusions can be drawn.

The marked differences in the frequencies of MONOGLOSS between Lam and Crosthwaite (2018) and other studies may be attributed to the methodology or genre effects. That is to say, there could be methodological factors that affect the observed frequency of MONOGLOSS in their analyses. This can include their interpretations of the MONOGLOSS category, their approaches to tagging MONOGLOSS versus heteroglossic tags, and units of analyses, among others. As the coding schemes were developed and adapted by the authors, the

interpretation of categories can influence the frequencies of identified engagement strategies. At the same time, the findings may also be because of differences in argumentative essays as opposed to other genres of academic writing. All in all, more evidence is needed to ascertain whether previous patterns (Fig. 4) are indeed due to the situational variables they examined, or mainly due to methodological choices. The potential benefits and drawbacks of previous studies are addressed in the following section (Section 2.5).

2.5 Benefits and Drawbacks of the Current Methodology to Investigate Engagement Strategies

There are benefits and limitations to current approaches to engagement resource analysis, which pertain to the intensive manual coding process (Lam & Crosthwaite, 2018; Lancaster, 2014; Wu, 2007; Xu & Nesi, 2019). One benefit of the current in-depth discourse analysis of engagement strategy is that it allows adaptation of the original Martin and White (2005) categories depending on the goals of the analysis and contextual factors. For example, although reasons are not explicitly mentioned in her paper, Ryshina-Pankova (2014) did not use the DENY category in her analysis of blog posts. Similarly, some researchers combine both Martin and White's (2005) categories and White's (2003) system, including the JUSTIFY category in their analysis (e.g., Lam & Crosthwaite, 2018). Detailed manual analyses also enable researchers to propose finer-grained categories that add more precision to the original ones that Martin and White (2005) can offer (Lancaster, 2014; S. H. Lee, 2017). In this regard, a bottom-up discourse analytic approach allows iterative refinement of the current Martin and White (2005) system to better capture evaluative meaning in academic discourse (see Hood, 2010; S. H. Lee, 2017).

Despite the major advantage of qualitative discourse analytic studies in terms of the theory-building process (Fuoli, 2018), one potential drawback of this approach is the analyses' scalability. As can be seen in Figure 2.4, the sample sizes in selected previous studies were as few as five (Ryshina-Pankova, 2014), and as many as 40 in Lam and Crosthwaite (2018). Two of the main contrastive studies between high- and low-graded essays to date (Lancaster, 2014; Wu, 2007) analyzed 30 essays at most. With this limited sample size, it is challenging to draw strong conclusions about patterns of engagement resource use across contexts. The limited sample sizes in a given study may also limit comparing the relative impacts of factors influencing the occurrence of engagement categories. For this reason, a new approach to the automated analysis of engagement strategies is clearly needed. Therefore, the current dissertation attempts to fill this gap in the literature by developing and evaluating a new automated system to analyze engagement strategies.

2.6 Summary

The current chapter has provided an overview of frameworks for investigating evaluative language in academic discourse. After reviewing three prominent theoretical frameworks (Biber et al., 1999; Biber & Finegan, 1988, 1989; Hyland, 2017; Martin & White, 2005), the chapter also justified using the engagement system in the Appraisal framework (Martin & White, 2005) as the theoretical framework for the present research. The chapter then described the entire systemic engagement network, comprising eight discrete strategies and one auxiliary strategy for the present dissertation. In the latter half of the chapter, I provided a synthesis of previous research investigating the distribution of engagement strategies across different populations. This synthesis indicated that while the engagement system may be helpful in uncovering certain

aspects of (epistemic) evaluative meaning in academic discourse, the qualitative nature of these methods may limit the generalizability and application of theoretical concepts in large-scale research and pedagogy. The overarching goal of the current dissertation, then, is to develop and evaluate a new probabilistic approach to identify and attach functional labels to evaluative language under the engagement system.

CHAPTER 3

STUDY 1: The Engagement Analyzer

3.1 Chapter introduction

The review of literature presented in Chapter 2 suggested that the system of engagement in the appraisal framework (Martin & White, 2005) may offer a useful theoretical framework to investigate epistemic evaluative meanings in academic discourse; however, the discourse-semantic orientation of this framework may have limited the scope of the methodological approach in previous studies. Specifically, the methodological aspects of previous studies may not warrant generalizable results because typical sample sizes ranged from 10 to 30 essays per study (e.g., Lancaster, 2014; Wu, 2007). Under the status quo, it would be challenging to make conclusive comments regarding the nature of academic discourse from the appraisal framework, although the case studies may still offer some important insights. This methodological challenge is the main driving force for the present dissertation study—that is, this dissertation attempts to contribute to the literature on evaluative language by developing an automated analysis tool that can conduct engagement resource analysis.

The three aims of Study 1 (reported in this chapter) are to:

- develop a discourse treebank of academic written English annotated for engagement resources (see Martin & White, 2005),
- train a series of machine learning models which can conduct engagement resource analyses of academic written English, and
- formally evaluate the performance of trained NLP models.

The present chapter is structured as follows. First, I discuss known challenges and important considerations in structuring corpus annotation projects on appraisal resources within

Systemic Functional Linguistics (Fuoli, 2018; Read & Carroll, 2012). Second, I provide an overview of the most recent version of the discourse treebank developed in the current dissertation—Engagement Discourse Treebank (EDT). In this part, I define the target domain of text (i.e., in-domain text) for this treebank; discuss both theoretical and practical considerations during corpus sampling; explain annotation layers (a total of four layers) and annotation schemes. Subsequently, the chapter presents a series of empirical studies on the training and evaluation of machine learning models. Finally, the chapter concludes with a few limitations of the current study and outlines possible future directions of discourse-oriented NLP in applied linguistics research.

3.1.1 Annotation of Appraisal Resources—Challenges and Current Practices

There seems to be a general sentiment in the literature that annotating appraisal resources or evaluative language by extension is challenging (Fuoli, 2018; Hunston, 2004; Read & Carroll, 2012). Researchers often describe evaluative language as elusive (Mauranen & Bondi, 2003), multifaceted/ layered (Thompson & Alba-Juez, 2014), and context-dependent (Hyland, 2005a; Martin & White, 2005). Evaluation, particularly the attitudinal aspect, is considered comparative, subjective, and value-laden (Thompson & Hunston, 2000). As a result, researchers have tended to focus more on the categorization of evaluative language, and thus there has been little discussion of the identification of potential evaluation from concrete discourse samples (Hunston, 2004).

A lack of explicit attention to the methodological aspects of identification and categorization appears to be the case in the literature on engagement, though with some exceptions (Fuoli, 2018; Read & Carroll, 2012). For example, only one of the five empirical

studies (Lam & Crosthwaite, 2018) reviewed in Chapter 2 reported actual numerical estimates of intercoder reliability (although their study also did not report the direct agreement between coders but rather the agreement of double-checks by two independently employed "reviewers"). Three other studies discussed general annotation procedures involving one or more researchers—how they calibrated the coding between two annotators and/or how they created context-specific manuals (e.g., Xu & Nesi, 2017); however, they tended not to discuss what changes were made from the original Martin and White (2005) framework and/or any additional considerations during annotation.

While the epistemological and methodological orientations of the previous studies (i.e., case studies that involve a great deal of qualitative discourse analysis) may make their research practices appear understandable and acceptable, the reliability of analysis (and transparency of the process) is the fundamental consideration even evaluated from the research standards proposed by researchers working within the postpositivist (Mackey & Gass, 2018), pragmatist (Creswell & Poth, 2016), and constructivist (Silverman, 2018) paradigms. For example, Silverman (2018) discusses inter- and intra-coder reliability as a prerequisite for a credible research report. Creswell and Poth (2016) stress that, although the process of intercoder agreement checks has often been neglected in qualitative research (p. 264), the systematicity and transparency of the process of coding are of paramount importance in qualitative inquiries. To remedy the situation, they proposed the following procedure for obtaining reliability in qualitative research:

Step 1: Establish a common platform for coding, and develop a preliminary code list;

Step 2: Define and share an initial codebook among coders;

Step 3: Apply the code book to additional transcripts, and compare coding across multiple researchers;

Step 4: Assess and report the intercoder agreement among researchers; and

Step 5: Revise and finalize the code book to inform further coding.

(Creswell & Poth, 2016, pp. 264–6)

Applying these steps in the research on engagement, it is apparent that the reported studies tend not to follow all five steps that ensure reliability in qualitative research. Although we can argue that each study started with a preliminary code list as they all draw on the same theoretical framework of the engagement system (Martin & White, 2005), they may not have employed multiple coders or reported how they modified the codebook according to their contextual needs. This challenge in the research practice of Appraisal analysis has been problematized recently by Fuoli (2018), who emphasizes reliability, replicability, and transparency in the annotation procedure for Appraisal analysis. The current study, therefore, follows the guidelines set by Fuoli (2018) in creating a gold-standard annotation dataset for engagement resource analysis. In the next section, we turn to this Fuoli's Stepwise annotation procedure.

3.1.2 Fuoli's Stepwise Annotation Procedure

The stepwise annotation procedure, proposed by Fuoli (2018), attempts to solve the problems in previous research practices in annotating Appraisal resources in the field. Acknowledging the challenges in achieving "perfect" agreement on discoursal or functional annotation, this approach emphasizes reliability, replicability, and transparency in the annotation process, which will, in turn, maximize the intercoder reliability of annotation. To this end, three principles are proposed:

Principle 1. All choices should be accounted for;

Principle 2. Annotation guidelines should be tested and refined until maximum reliability is achieved;

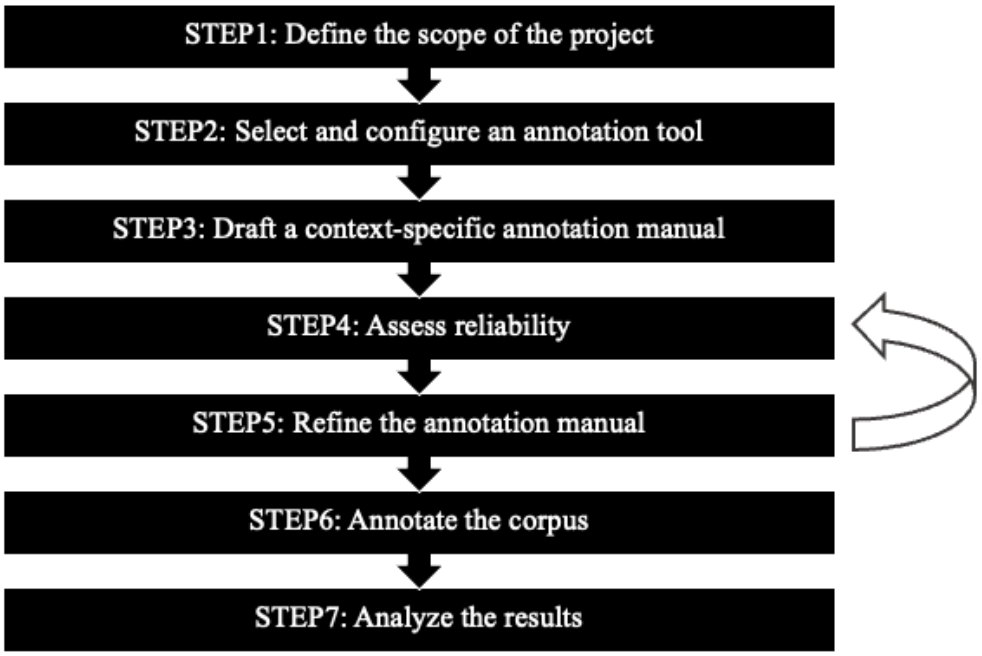
Principle 3. Reliability should always be assessed, and reliability scores reported and discussed. (Fuoli, 2018, pp. 246–7)

During an annotation project, it is likely that annotators will encounter instances that are not prototypical of the categorical description of one category. For example, Fuoli (2018) notes that the lexical verb *believe* can be tagged as ENTERTAIN or PRONOUNCE in a given context, and this distinction can be tricky. Following Principle 1, the annotation team should strive to document the decisions they make during annotation and discuss possible alternative interpretations of specific cases as a team. Principle 2 states that these discussions should inform the refinement of annotation guidelines progressively until both general and context-specific issues are saturated. At this point, the annotation team should also progressively track the intercoder reliability of annotation and see if refinement of the guidelines improves the reliability of annotations. Like Principle 3, Fuoli (2018) emphasizes the importance of reporting intercoder reliability estimates. The present study follows these three general principles proposed by Fuoli (2018) in annotating engagement resources.

Figure 3.1 presents the seven-step procedure recommended by Fuoli (2018). For the sake of simplicity, not all steps are described here (see Fuoli, 2018 for details); however, it is important to highlight how the overall procedure contributes to greater transparency, replicability, and reliability. Steps 1–3 concern the initial part of the annotation project, where a researcher defines the scope of the project (including the genres of texts and other important contextual features), whether it is quantitative or qualitative in nature (Fuoli, 2018). At this stage,

a context-specific annotation scheme should include definitions of each annotation category, some examples from the same domain of texts, and how instances can be annotated; as well as things to consider during annotation (i.e., flowcharts). Given a context-specific annotation scheme, the annotation team should conduct a series of trial rounds (Steps 4 and 5), where the initial coding scheme is applied to their data and inter-annotator reliability is progressively tracked. The feedback loop from Step 5 to Step 4 is critical at this stage because it ensures that annotation manuals are optimized for the purposes of a given annotation project. Finally, Step 7 ensures that the reliability of the entire annotation project (the entire steps) is reported and progressively informs subsequent research in the same domain. In the present study, these three principles and a stepwise annotation procedure are applied to develop a treebank for engagement resource analysis.

Figure 3.1
Fuoli's (2018) stepwise annotation procedure



3.2 This Study

The goals of the present study are twofold. First, it aims to create a treebank for engagement resource analysis in academic written English, drawing on the engagement system in the appraisal framework (Martin & White, 2005). Second, it aims to train an end-to-end machine-learning model which performs automatic analysis of engagement resources in the same domain. The study is guided by the following three research questions.

3.2.1 Research questions

1. What are the levels of intercoder agreement after the annotators with linguistics backgrounds are trained on adapted schemes of the engagement system?
2. What are the impacts of machine-learning architecture selection and associated hyperparameters on precision, recall, and F1 scores?
3. What are the precision, recall, and F1 scores of the best-performing pipeline of the Engagement Analyzer?

3.3 Method

3.3.1 Engagement Discourse Treebank (EDT)

3.3.1.1 Definition of in-domain text

In the NLP literature, the notion of in-domain text is often used to talk about the domain of a text on which a given NLP component is trained to perform a certain task (Ramponi & Plank, 2020). Presumably, ML models will perform best when the model is used to make inferences for the type of texts they were trained on. Importantly, the logic here is analogous to the notion of the target language use (TLU) domain in language assessment (e.g., Bachman & Palmer, 2010) in that both state that tools—whether it is for NLP or assessment—should be

constructed with features of the language of the target domain in mind. Since the purpose of the Engagement Analyzer is to analyze interpersonal rhetorical strategies in academic written English, I define the in-domain text of the Engagement Discourse Treebank (EDT) broadly as academic written English in varying genres written by L1 and L2 writers.

3.3.1.2 Corpus sampling approach

Based on the definition of in-domain text—academic written English in varying genres written by L1 and L2 writers—I chose to sample a wide variety of texts from corpora that closely align with this definition. Curating texts from corpora that closely match the TLU domain in terms of its text production features (e.g., communicative purposes, writers) can support the argument for the applicability of the resulting NLP tool to analyze the interpersonal features of academic written English. Table 3.11 lists the five corpora I sampled the data from, when constructing the EDT. It describes important situational information about the texts included. While these corpora match the definition of in-domain text presented above, they also cover a wide range of text variations within the domain, particularly in terms of writers' backgrounds, writers' proficiency levels, and genres of writing (e.g., essay, argumentative writing, lab report). In what follows, I briefly describe each of the source corpora.

Table 3.1
Source corpora of the EDT.

Corpus	General textual feature	Genres	Language background	Estimated proficiency levels	Target proportion
BAWE	UK univ. assignments	Varying	L1 + L2	C1 +	0.35
MICUSP	US univ. assignments	Varying	L1 + L2	C1 +	0.35
ICNALE	Timed essay	Argumentative	L2	A2 – B2	0.1
TOEFL 11	Exam response	Argumentative	L2	B1 – C1	0.1
FCE	Exam response	Argumentative/ letter writing	L2	B2	0.1

British Academic Written English (BAWE)

British Academic Written English (BAWE) is a 6.5-million-word corpus of authentic university writing, collected in several UK-based universities in 2006 and 2007 (Alsop & Nesi, 2009; Nesi & Gardner, 2012, 2018). It includes a total of 2,761 complete written assignments submitted to these universities for degree-obtaining purposes by both undergraduate (freshman to senior) and post-graduate (MA-level) students from various disciplines. BAWE is one of the most representative corpora of academic writing in that it samples written assignments from more than 30 specific disciplines across four different broad subject areas: Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences. The BAWE corpus also provides rich meta-data as to course title, degree, and writer's L1 background, as well as specific genres of writing (closely classified by the researchers internally). A total of 1,953 papers were submitted by L1 writers, and 808 papers were written by L2.

Michigan Corpus of Upper-level Student Papers (MICUSP)

MICUSP is another corpus of authentic university written assignments, documenting 829 papers submitted as course assignments to a university in the United States (Römer & O'Donnell, 2011; Römer & Swales, 2010). It includes writing samples from a total of 16 disciplines. A total of 681 papers were written by L1 writers, and 148 papers were submissions by L2 writers. MICUSP classifies the submitted assignments into seven genres or assignment types.

TOEFL 11

TOEFL 11 (Blanchard et al., 2013) is a collection of 12,100 timed argumentative essays written by non-native writers of English as a part of the TOEFL® exam in 2006–2007. The collected writing samples were based on eight different independent writing prompts from TOEFL, which are no longer used in the current test administration. The corpus represents writers speaking 11 different first languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. Texts from this corpus are characterized as timed argumentative writing, which is a kind of writing task often used within standardized English proficiency tests for university admission purposes (thus closely matching the definition of in-domain text above). Presumably, the current dataset mostly represents L2 writers with CEFR levels of B1 to C1.

CLC FCE Dataset

The Cambridge Learner Corpus (CLC) of First Certificate in English (FCE) contains 1,244 exam scripts, written by test-takers of the Cambridge ESOL FCE (Yannakoudakis et al.,

2011). One administration of the CLC FCE writing section includes two different tasks, which are chosen from the following text genres: a letter, a report, an article, a composition, or a short story. Although the assumed proficiency range was narrow compared to TOEFL 11, the current corpus allowed wider representation of discourse features that are used in L2 proficiency exams. This corpus has been used previously to construct a syntactic dependency treebank of L2 writing (Berzak et al., 2016). The corpus is available for non-commercial research and education purposes.

International Corpus Network of Asian Learners of English (ICNALE)

The ICNALE corpus (Ishikawa, 2013, 2018) is a collection of timed argumentative essays written by L2 writers with Asian backgrounds. The corpus elicited written responses for research purposes, thus slightly lacking in terms of authenticity of language use. All research participants (mostly university students in Asian regions recruited from the project team) completed two essays with the same sets of prompts during the data collection (Part-Time Job and Smoking in restaurants). Although the corpus is narrow in terms of the topics, it was used to augment the rhetorical strategies of L2 writers with lower language proficiency (roughly from CEFR levels A2 to B2).

3.3.1.3 A minimal context window approach in corpus sampling

During corpus sampling, I opted for a minimal context window strategy to strike a compromise between the validity of the following annotation and any practical considerations. Specifically, I chose to employ a window of three contiguous sentential segments as the unit of

analysis for annotation data. This approach has its pros and cons, some of which are practical, such as the copyright of the source corpus and the cost of annotating it.

In an ideal situation, the unit of analysis for annotation should be an entire document, mainly because the object of the annotation is discoursal; however, this was not possible due to, for example, potential copyright issues with some of the corpora used in the study (if the treebank is to be made available to the public). For example, although the BAWE corpus is open access under a Creative Commons license, corpora such as ICNALE and CLC FCE are distributed under stricter licenses. This limited the current annotation project to operating within the fair use limit. For example, the CLC FCE states that not more than 200 running tokens should be disclosed.

One crucial decision point in determining the unit of analysis included a practical consideration concerning the extent to which the resulting annotated corpus can represent diversity in rhetorical devices, given the limited resource. For example, annotating the entire 6-million-word BAWE corpus within a 3-month timeframe would be impractical with limited funds. One advantage of the current three-sentence window approach is that the dataset can include a wider range of writing styles and stances compared to using the whole document as the unit of analysis. Assuming that a writer weaves a specific stance throughout a document, using the entire document as a unit of analysis may skew the dataset unless the corpus represents enough documents (probably more than 200 writing samples). As it turned out that the size of the most recent version of EDT is around 4,000 sentences, the annotation of 200 writing samples would have been beyond reach. Using a contextual window approach allowed representing a broader range of rhetorical and style features, while not discarding the contextual information necessary to evaluate the rhetorical features of a given lexico-grammatical choice.

Another practical reason for the minimal context approach is the fact that commonly used Transformer models (e.g., BERT, RoBERTa; described below in sections 3.2.2.3) only take 512 (sub-word) tokens at once (although there are variants of Transformer models that double these sizes, they are not considered here). Consequently, even if the whole text is annotated, the machine learning components take a moving window approach during training and cannot take full advantage of contextual information, at least as of 2022–3.

While the aforementioned reasons motivated the decision to use a minimal context approach, there are certainly some caveats to this approach. One important caveat is omitting potentially critical discourse moves that precede or follow a particular context window. In such cases, neither annotator nor ML models will be able to consider that information in the current approach. While this may affect the validity of annotation, the minimal contextual approach is still believed a viable first step toward discourse-oriented NLP. Further drawbacks are discussed in the limitations section of this study, which also makes recommendations for future research.

3.3.1.4 Corpus annotation scheme

The Engagement Discourse Treebank has a total of four layers: (a) Clause boundary, (b) Engagement span and category, (c) Engagement hierarchy, and (d) Supplementary rhetorical move layer. The purposes and specific tags of each layer are described below.

Clause boundary layer

The first layer of the EDT is clausal boundary annotation and is structural in nature. This layer was introduced to help annotators grasp the general syntactic structure, such as that-clause complementation, of each annotation sentence. Five tags were selected for this layer: T-unit

(Hunt, 1965), Main clause, Subordinate clause, Embedded clause, and Fragment. Table 3.2 provides definitions of each tag. See Figure 3.2 (page 92) for an illustrative example of this layer.

Table 3.2
Clausal boundary layer tags

Clausal layer label	Definition
T-unit	A T-unit consists of a MAIN clause and any dependent (i.e., SUBORDINATE or EMBEDDED) clauses attached to it.
MAIN	An independent clause, which functions as a complete unit.
SUBORDINATE	A dependent clause attached to a main clause through the use of subordinate conjunctions (e.g., because, although, if, when, as, while, etc.)
EMBEDDED	A type of dependent clause that functions as a part of another clause. That is, an embedded clause is included in a subject, object of another clause (i.e., complement clause) or functions as an adjective to modify a noun (i.e., relative clause). Inserted clauses, such as parataxis, are also considered a type of embedded clause.
FRAGMENT	An incomplete sentential unit, typically, without any verbs

A few notes are required in order to explain the purposes behind each tag. T-unit (Hunt, 1965) is a frequently used unit of analysis in second-language writing research (Lu, 2011; Norris & Ortega, 2009), it is often used in research on the engagement system (see Ryshina-Pankova, 2014). Utilizing information from this tag, it is possible to characterize the overall engagement strategy in each T-unit, which may be helpful, depending on the purpose of the research. The MAIN, SUBORDINATE, and EMBEDDED clause tags were particularly useful in annotation. For example, the annotators were able to focus on identifying any Engagement resources in the MAIN clause first, which tend to determine the overall tone of sentences before shifting their focus onto minor rhetorical moves which do not necessarily reflect writers' overall stances (see also Engagement hierarchy layer).

Engagement category layer

The engagement category layer is the primary focus of the annotation project. The annotators were asked to do two tasks for this layer—identifying the span of the text which encodes any of the engagement meaning, and providing an exact label for the span. To this end, eight categories were adapted from the original engagement system by Martin and White (2005), and the category descriptions were adapted from previous studies (Wu, 2007; Xu, 2020). Table 3.3 lists engagement strategy labels (second column), label descriptions (the third column), and overarching discourse strategies to which each engagement strategy belongs (first column). For complete guidelines, see guidelines online; Figure 3.3 on page 92 illustrates this layer.

Engagement hierarchy layer

During the annotation project, the annotation team noticed that engagement strategies could sometimes occur inside the presupposed content of an utterance. This added extra details about how writers position themselves. A review of previous research (and reported excerpts; Chang & Schleppegrell, 2011; Hood, 2010; Lancaster, 2014; Martin & White, 2005; Ryshina-Pankova, 2014; White, 2003; Wu, 2007; Xu & Nesi, 2017) indicated that there was little discussion in the literature on how to deal with these "secondary" engagement strategies (here, the term secondary is used to indicate that these may not characterize the entire engagement strategy of an overall T-unit or clause).[2] Nevertheless, to understand the precise stances that writers adopt in discourse, this information about primary/ secondary engagement can be helpful. For this reason, a binary classification was made for each engagement tag as to whether it affects the overall engagement value of the T-unit or only a subset of propositions within the T-unit.

Supplementary rhetorical move layer

[2] This lack of explicit guidelines is in line with Fuoli's (2018) claim for the issue of transparency in Appraisal analysis.

Table 3.3
Engagement category layer tags (Adapted from Wu, 2007; Xu, 2020).

Macro strategy	Engagement strategy	Description	Examples of prototypical lexico-grammatical realizations (see Chang & Schleppegrell, 2011)
Contraction	Disclaim: Deny	An utterance which invokes a contrary position but which at the same time rejects it directly. The contrary position is hence given very little dialogic space.	• Negative particles (e.g., *not, never*)
Contraction	Disclaim: Counter	An utterance which expresses the present proposition as replacing and thus 'countering' another proposition which would have been expected.	• Conjunctions (e.g., *but*) • Adverbials (e.g., *however*) • Adverbial clauses (e.g., *although*)
Contraction	Proclaim: Concur	An utterance which shows a writer's expectation/ assumption that putative readers will agree with the proposition and/or have the same knowledge.	• Adverbials (e.g., *indeed, of course*) • Adverbial clauses (e.g., *as one expects*) • Display questions; tag questions
Contraction	Proclaim: Pronounce	An utterance which expresses a strong level of writer commitment through the author's explicit emphasis and interpolation, thereby closing down the dialogic space.	• Mental/Communication verbs (e.g., I *contend*, we *conclude*, I *propose*) • Emphatic do (e.g., I *do* believe that) • Modal attributes (e.g., *it is evident* that)
Contraction	Proclaim: Endorse	An utterance which refers to external sources as warrantable, undeniable, and/or reliable. It expresses the writer's alignment with and endorsement of an attributed proposition. As such, the dialogic space is somewhat narrowed.	• Reporting verb (e.g., Kyle (2020) *demonstrated* that)

Table 3.3 (Cont'd)

Macro strategy	Engagement strategy	Description	Examples of prototypical lexico-grammatical realizations (see Chang & Schleppegrell, 2011)
Expansion	Entertain	An utterance which indicates the author's position but as only one possibility amongst others, thereby opening up dialogic space.	• Modal verbs (mainly epistemic and deontic modals; e.g., *may*, *would*; Palmer, 2001) • Mental/Communication verbs (e.g., I *think*/*suppose*, we *suggest*) • Adverbials (e.g., *perhaps*, *probably*) • Adverbial clauses (e.g., *unless*, *when*, *if*) • Modal attributes (e.g., *it is likely that*) • Evidentials (e.g., *seem*, *apparently*)
Expansion	Attribute	An utterance which signifies dialogic space as the writer attributes the proposition to an external source.	• Adverbials (e.g., *reportedly*) • Reporting verbs (e.g., *they argue*/ *believe*)
Monogloss	Monogloss	An utterance which does not employ any value of engagement. Such an utterance ignores the dialogic potential in an utterance.	• Present-tense verbs • Lacks of any other engagement strategies
Auxiliary	Justify	An utterance which engages in persuasion through justification or substantiation.	• Adverbials (e.g., *therefore*, *for this reason*) • Conjunctions (e.g., *because*)

Note. Adapted from Martin and White (2005), Wu (2007), and Xu (2020).

While the primary focus of the EDT is to annotate academic texts in terms of interpersonal meaning-making resources under the engagement system (Martin & White, 2005), it is also important to acknowledge that there is a considerable degree of interaction between interpersonal and textual meta-discourse in academic language use (Halliday & Matthiessen, 2014; Hyland, 2005a; Thompson & Hunston, 2000). During the initial annotation phase, the annotation team noticed that a clause could be framed under MONOGLOSS not only due to the writer's bare assertion but also its textual function in discourse. For example, a textual segment can make an intra-textual reference (e.g., *Table XX summarizes the result of the experiment*) or signal intra-textual organization (e.g., *The paper is organized as follows*). This observation suggests that not all MONOGLOSSIC statements should have equal value in discourse. In the original Martin and White (2005) scheme, they presented qualitative analyses of newspaper articles, not academic papers, so they might not have encountered this challenge in classifying MONOGLOSS into different types.

For this reason, we decided to supplement the Engagement labels with additional rhetorical moves that could account for such interaction between interpersonal and textual discourse. Many of these tags were adapted from the meta-discourse literature (Interactive resources in the meta-discourse literature, Hyland, 2005), previous studies on academic source use (e.g., Nesi, 2021), and an SFL-based discourse analytic framework (e.g., Martin & Rose, 2007). Three major functions were covered with 12 specific discourse tags. Table 3.4 lists these supplementary tags available in the EDT. Figure 3.4 illustrates this layer's annotation.

Table 3.4
List of supplementary tags annotated in EDT.

Category	Supplementary tags	Description	Examples
Reference to internal or external information			
	a) Citations	Mention external source(s) in the text in the form of in-text or narrative citation.	(Martin & White, 2005); Ortega (2009)
	b) Sources	Mention external source(s) in the text in the form of nominal expressions.	Annual report by X; numerous studies
	c) Endophoric markers	Refer to information in other parts of its own text	as noted above, see Fig. 1, in section 2
	d) Quotes	Segments of text with direct quotations (including quotation marks, both single and double)	"Stay hungry. Stay foolish."
Logical connections			
	e) Exemplifying	Signal illustrations/examples in the text (mostly internal).	for example, to illustrate, e.g.
	f) Expository	Signals elaboration/clarification in a subsequent part of a text (mostly internal).	in other words, that is, i.e., I mean, this means
	g) Additive-internal	Signals additional argumentative elements; focuses only on internal relations.	additionally, moreover, besides
	h) Comparative-internal	Marks an argument as similar or different. This category has to be differentiated from a COUNTER move.	similarly, by contrast, conversely

i) Justifying	Signals persuasion through justification or substantiation. For reasons of simplicity, we include both external and internal relations.	because of X, due to X, therefore
Text organizing devices		
j) Goal-announcing	Signals the purpose/ goals of the text itself	my purpose is, Section 2 describes, the chapter/ section focuses/ proposes, we intend to, in this chapter
k) Text-sequencing/ staging	Signals sequences and stages of argumentative elements in the text	First*, Lastly*, to start with, so far, overall
l) Summative	Signals summary/ conclusion of part of the text	to conclude, in short, to sum up, the/our conclusion is

Figure 3.2
Illustration of the clause boundary layer (batch2_1452).

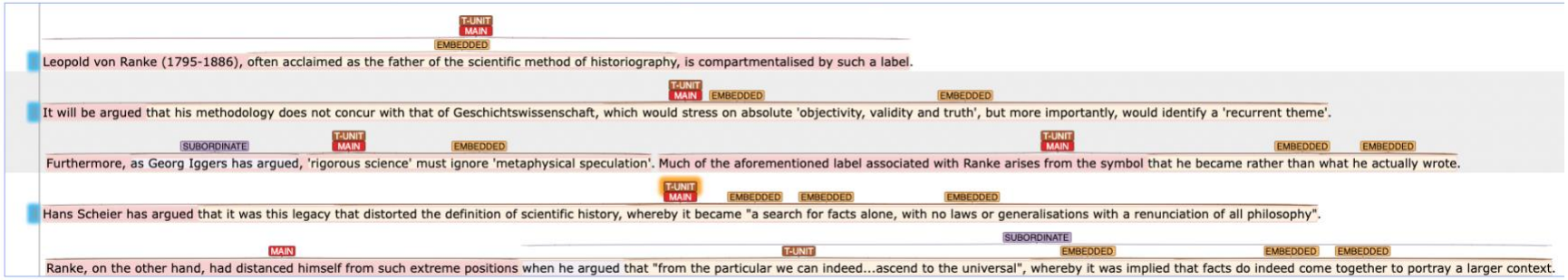


Figure 3.3
Illustration of the Engagement annotation layer (batch2_1452).

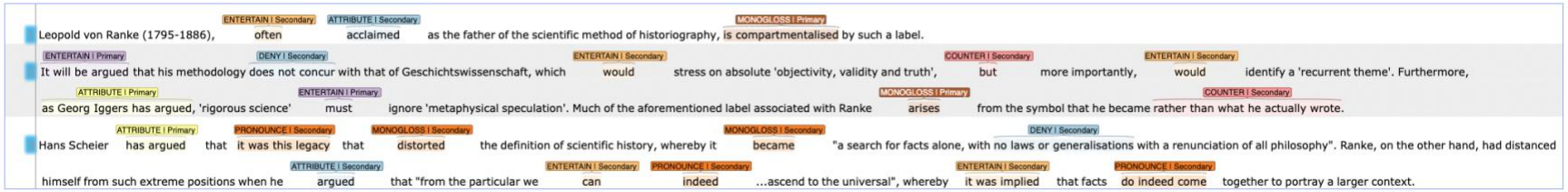
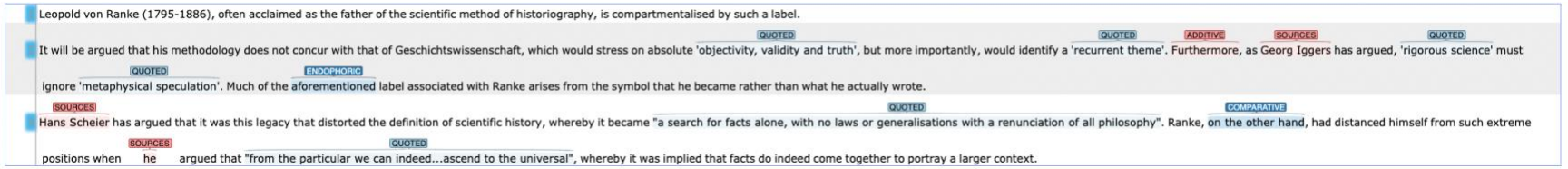


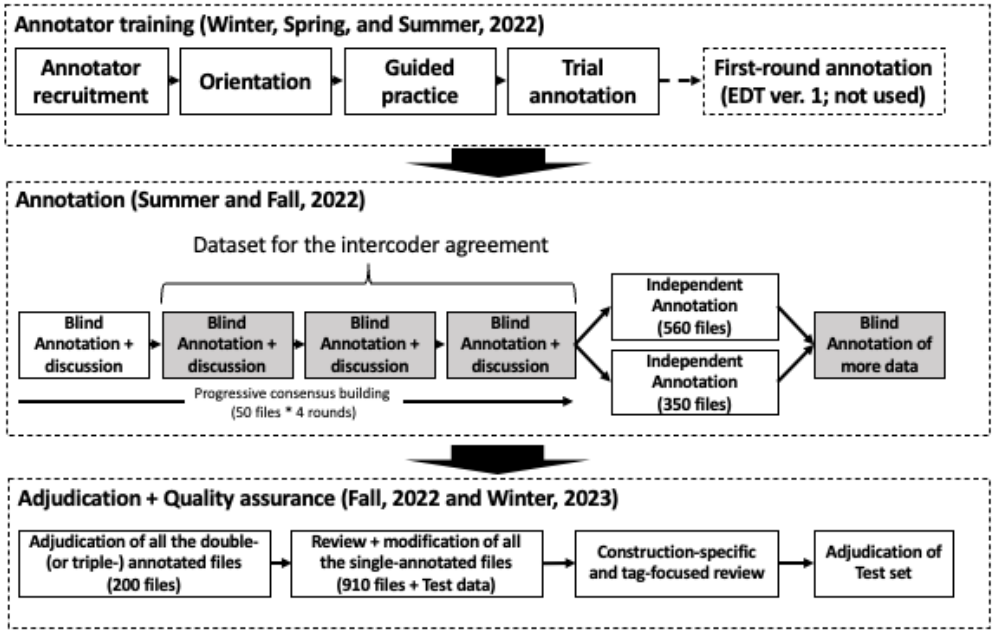
Figure 3.4
Illustration of the supplementary rhetorical move layer (batch2_1452).



3.3.1.5 Annotation procedure

Figure 3.5 presents an overview of the entire annotation procedure of the EDT. A three-stage procedure was followed to ensure annotation quality: a) annotator training, b) annotation, and c) adjudication and quality assurance. The goals and detailed process of each stage and their sub-phases are described below.

Figure 3.5
Overview of the annotation procedure of the EDT.



3.3.1.6 Annotator profiles and training

Two annotators were recruited from an upper-division linguistics course. They were both first-language speakers of English, and they had completed introductory linguistic courses, which covered syntax and semantics. It is important to note that although these annotators were introduced to a usage-based, functional approach to linguistic structures, neither of them had taken SFL-based functional linguistics courses; thus, a training scheme was developed to

introduce fundamental concepts in Systemic Functional Linguistics (primarily about three meta-functions of language) as well as the Appraisal framework. Extensive hands-on training procedures were developed. The training procedure included: orientation, guided practice, and trial annotation with discussion. The training took place in the Winter and Spring terms of 2022.

Orientation phase

The first phase of annotator training was orientation. During this phase, the annotators were introduced to the basic concepts of SFL and the Engagement system, including the distinction between monoglossia and heteroglossia, the distinction between contraction and expansion, and the notion of attribution (see https://egumasa.github.io/engagement-annotation-project/3_Categories/). Also during this phase, preliminary topics in lexico-grammatical analysis were introduced, such as constituency, finite and non-finite clauses, subordinate clause, embedded clause, and T-unit (https://egumasa.github.io/engagement-annotation-project/1_Basic_grammar/).

Guided practice phase

During this phase, the annotators went through multiple-stage practice with iterative feedback guidance from the researcher. First, they were introduced to the annotation tool, WebAnno version 3.2 (Eckart de Castilho et al., 2016; Yimam et al., 2013). WebAnno was used throughout the annotation project as the graphical user interface. Second, a sample of 500 sentences was distributed to the annotators. They annotated this training sample independently, which was later checked for agreement with the researcher's annotation (which was treated as a gold tag at this point). For each annotator, the researcher identified patterns of errors in the training, provided tailored feedback independently, and clarified any concepts in the guidelines. The annotators were then asked to repeat their annotation of the same training sample to

implement their understanding. This stage was intended to give the annotators an opportunity to familiarize themselves with the concepts and meta-language to talk about the Engagement system with concrete examples.

Trial annotation with discussion

During the final phase of annotator training, the annotators were introduced to another set of 200 sentences and annotated this sample independently. Blind annotations of the two annotators were then compared by the researcher. At this point, weekly meetings were held to discuss any inconsistencies and potential resolution strategies to be added to the guidelines. This process was repeated with another set of 200 sentences, where the two annotators and the researcher established a shared understanding of concepts of the engagement system. This process ensured that all three annotators (including the researcher) had seen various realizations of engagement meaning in in-domain data. The annotators were assigned to engagement annotation tasks for different subsets of the corpus during summer 2022, but these files were treated as version 1 files of the EDT and, therefore, not included in the most recent version of EDT.

Annotation process

The data annotation took place in three phases in the Summer and Fall terms of 2022. First, to ensure a higher level of consistency and measure intercoder agreement, both annotators and the researcher blindly annotated the first 50 files from the corpus (comprising approximately 125 sentences). The three versions were compared side-by-side during a meeting to resolve inconsistencies in annotation and discuss necessary changes to the guidelines. I made any necessary adjustments to the annotation, and the resulting adjudicated files were added to the annotated corpus.

After the first batch, both annotators blindly annotated the same new mini-batch of data (each with 50 files), and intercoder agreement was tracked progressively. After every 50 files, meetings were held to resolve any issues with the new samples, and revisions were made to the tagging guidelines (Steps 4 and 5 in Fuoli's method; Fuoli, 2018). This additional consensus-building process was repeated three times, totaling 150 files independently coded by the annotators, followed by close discussion. The two versions of first-pass annotations for these 150 files were used to calculate intercoder agreement and used as a human-level annotation benchmark. Overall intercoder agreement at this stage showed that the annotation was a relatively difficult task, considering Cohen's kappa of .67 and macro F1 of .63 (using one of the annotations as a reference for the F1 measure). More details of intercoder agreement are discussed in Section 3.4.1, Intercoder agreement.

Once a human-level agreement benchmark was established, the two coders were assigned to different mini-batches to speed up the annotation progress. At least one annotator tagged each file during this independent annotation. One of the annotators completed approximately 560 files, while the other completed 350 files at this stage. Before the files were added to the final dataset, they were all reviewed and modified as necessary by the researcher. Once the annotators completed all the assigned files, they were assigned to annotate additional data, which were shared between them.

3.3.1.7 Quality assurance

Given the relatively low intercoder reliability of the initial annotation, additional steps were taken to ensure annotation consistency throughout the corpus. First, every single annotation file was reviewed and modified by the researcher. Other strategies were used to enhance the

consistency of annotation, including (a) building annotated corpus databases, (b) indexing problematic examples, (c) a focused review of problematic examples, and (d) post hoc tag fixes through the concordance tool.

Annotated corpus databases.

Fuoli's first principle says that all decisions should be accounted for (Fuoli, 2018). Throughout the annotation project, each annotation file was stored as an entry in a database, where all members of the annotation team were able to document any annotation decisions. The database documents not only meta-data, such as file names from the original corpus and the status of annotation, but also annotators' questions and any discussion of specific cases.

Indexing problematic examples

Another advantage of the database also included the ability to index any annotator-generated issues in the database as an attribute or "problematic case". Using this linked database, the annotation team documented any problematic cases to be resolved later in the adjudication stage. For example, when an annotator noticed any inconsistency in the annotation guidelines, such as the annotation span related to existential construction + negated nominal phrases (e.g., "there is no reason", "there is no doubt"), they were encouraged to create a new tag for this and index the annotation example in the database. Such indexing allowed the team to conduct comprehensive retrievals of any problematic cases throughout the corpus for discussion and review of the data for consistency.

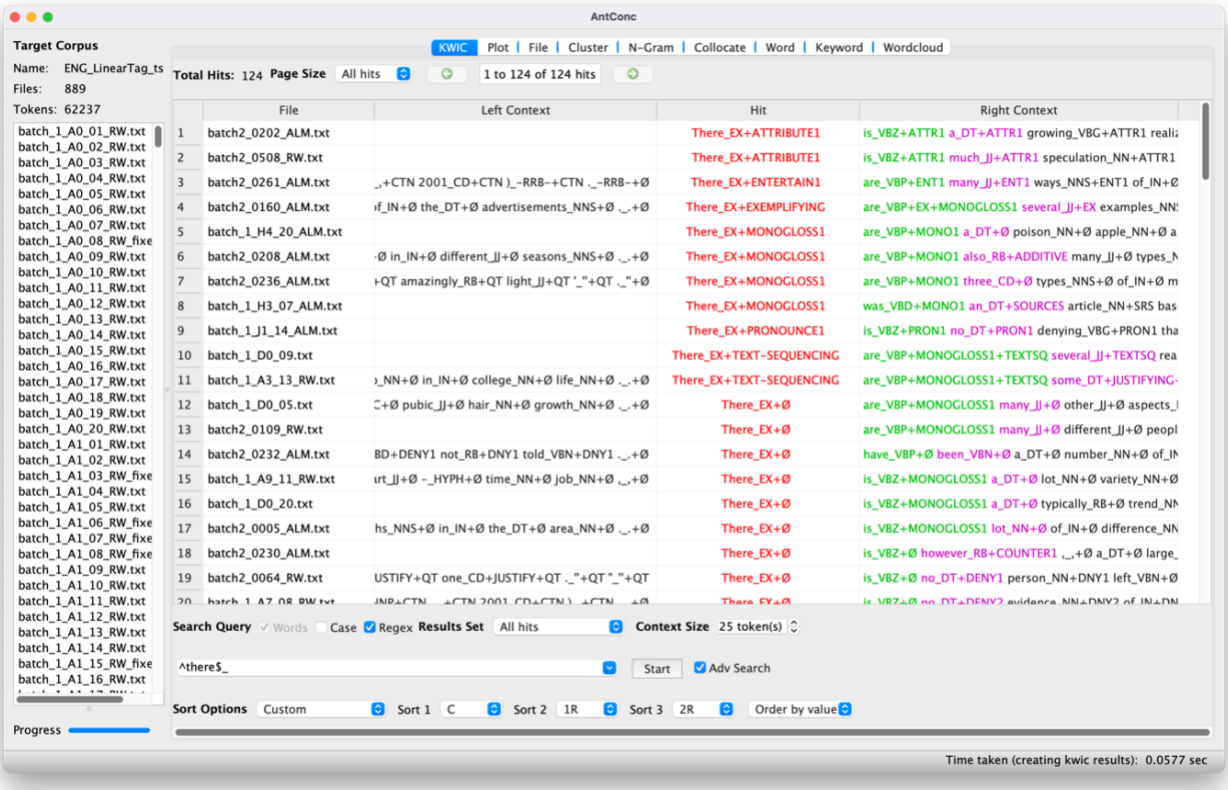
Focused review of the entire annotation using concordances

A focused review of problematic cases was conducted to ensure consistency in the annotation throughout the corpus. Based on descriptions of issues arising in "problematic cases" entry, comprehensive queries were made throughout the corpus, using concordance software

AntConc version 4 (Anthony, 2021). To be able to search for cases through concordances, the CONLL-like vertical representation was converted to a linear text format separated by underscores delimiting the tokens and sequences of tags. Each tag field was formatted as Penn POS + alphabetically sorted Engagement tags.

Using this linear representation of annotated data, concordance searches were conducted to retrieve cases to review. Figure 3.6 illustrates this process, where the search term "^there$_" returned all the occurrences of there in the annotated data. As can be seen in the figure, a total of 124 occurrences were retrieved, which were sorted according to the tag associated with the search term. Among the 19 cases visible in Figure 3.6, a variety of tags were attached to existential constructions (as can be seen from the EX tag in the POS field)—two ATTRIBUTE, one ENTERTAIN, one EXEMPLIFYING, four MONOGLOSS, one PRONOUNCE, two TEXT-SEQUENCING, and eight empty tags (zeros). Although the sole variety of tags is not a problem here—indeed, we identified that existential construction can be used to ATTRIBUTE (e.g., line 2: "there is much speculation") or PRONOUNCE (e.g., line 9, "there are no denying")—we identified that there were inconsistencies in our span of MONOGLOSS. For instance, lines 5–8 included existential "there" in the span of MONOGLOSS, while it was excluded in other cases (e.g., lines 12, 13, 15–17). During the adjudication, we focused on such inconsistencies and fixed them in our annotation. In this case, we clarified the span of MONOGLOSS in existential clauses (i.e., MONOGLOSS on a linking verb, excluding existential there), and fixed four files from lines 5–8 (see File ID in the File column).

Figure 3.6
Illustrative AntConc display for tag-specific adjudication.

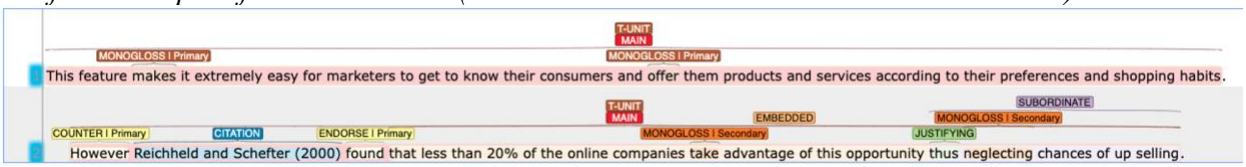


3.3.1.8 Sample Data

In this last section about the EDT, I will present two excerpts from the corpus to further illustrate the nature of our annotation. The first example (Figure 3.7) comes from the BAWE corpus (Writing ID: 3081c; Sentence 2 of the 33rd paragraph), which is a critique submitted to the Hospitality, Leisure & Tourism Management department about the travel trade.

Figure 3.7

The first example of annotated data (MONOGLOSS >> COUNTER >> ENDORSE).

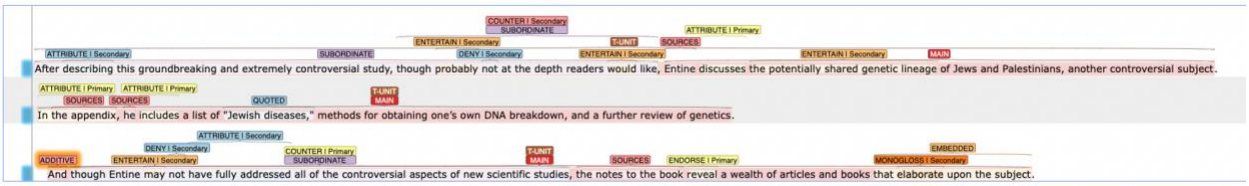


Note. EDT id = Batch2 1238.

At the schematic level, this excerpt shows a combination of monoglossia (first line) and heteroglossia (second line). In the first sentence, the writer makes assertions about a previously mentioned "feature". Here, the writer is presenting their self-sourced argument or evaluation without presenting any possible alternatives (hence MONOGLOSS). In the second sentence, the writer starts by COUNTER-ing the previous discourse and then provides some evidence by referring to an academic source (i.e., Reichheld and Schefter, 2000; thus CITATION). The writer does this by also showing some level of commitment to the attributed information (e.g., *found*), rather than only reporting it, thus this is considered ENDORSE, instead of ATTRIBUTE. It is also important to note, in this example, that the expression "according to" in the first sentence was not tagged as ATTRIBUTE considering the rhetorical function for which it was used (i.e., considered equivalent to "depending on" rather than attribution to a specific viewpoint). In existing corpus-based approaches to the analysis of epistemic stance-taking features, which use regular expressions to identify these items (e.g., Yoon, 2017b), these functional implications of an item would not be considered.

The second example (Figure 3.8) was taken from MICUSP (writing ID = POL.G0.25.1; Sentence 3 of paragraph 27). In this excerpt, the writer makes attribution to external sources throughout, but with increasing alignment to attributed information.

Figure 3.8
The second example of annotated data (ATTRIBUTE >> ATTRIBUTE >> ENDORSE).



In the first sentence, the overall writer's rhetorical move is to ATTRIBUTE (e.g., After *describing* … Entine *discusses* …); thus, it was considered as primary engagement. This is accompanied by some secondary engagement strategies such as COUNTER (e.g., *though* […] not at the depth readers would like) and ENTERTAIN (e.g., the *potentially* shared genetic lineage). This overall stance continues in the next sentence, where the writer refers to Entine's work (e.g., i*n the appendix* he *includes*). In this example, the phrase (*in the appendix*), which is often treated as an ENDOPHORIC references in previous corpus tools (Bax et al., 2019), was interpreted as an external source (belonging to Entine's work not the immediate writer's). In line 3, this general attributive stance is further reinforced by additional mention of a SOURCE (i.e., "the note to the book") with additional ENDORSEment of the content of it (i.e., *reveal*).

3.3.1.9 Interim Summary

To summarize, the Engagement Discourse Treebank (EDT) is a fully human-annotated corpus of academic English writing based on the adapted framework of the engagement system (Martin & White, 2005). It also provides a systematic tagging scheme for corpus-based discourse analysis, which can standardize the annotation process in research on the Appraisal framework (Martin & White, 2005). Further, the supplementary rhetorical moves, inspired by the research on metadiscourse (Hyland, 2005a) and academic source attribution (Nesi, 2021), allow us to take into account possible interactions between interpersonal and textual discourse functions, which

the engagement system alone cannot capture. These items are all tagged based on the functions they accomplish in the discourse, not based on the surface structure, allowing the possibility of developing automated discourse-oriented NLP tools that go beyond the traditional regular-expression-based tools often used in second language research (e.g., Bax et al., 2019; Yoon, 2017b). The next section discusses the machine learning architecture to accomplish the automation of this discourse annotation goal using state-of-the-art neural network NLP models.

3.3.2 The Engagement Analyzer

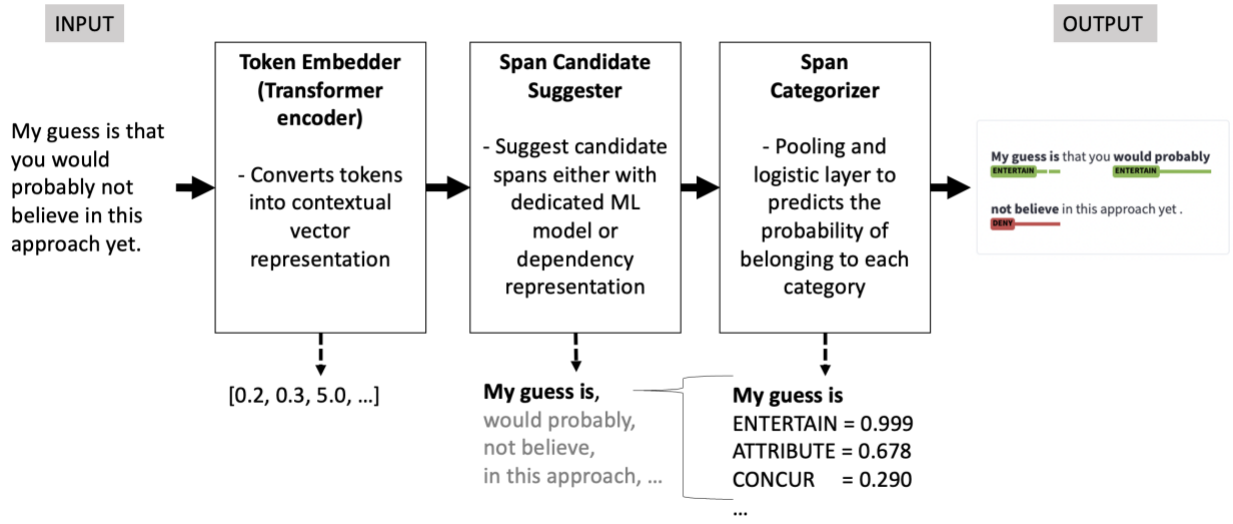
3.3.2.1 Overview

The primary goal of Study 1 is to develop and evaluate a natural language processing pipeline that can conduct engagement resource analysis (Martin & White, 2005). To this end, we constructed the Engagement Discourse Treebank (EDT) introduced in section 3.3.1. The current section details the design of the end-to-end machine learning component of the Engagement Analyzer. Machine learning models were implemented using the spaCy Python package (Honnibal et al., 2020) and the HuggingFace library (Wolf et al., 2020).

3.3.2.2 Natural language processing (NLP) pipeline

The natural language processing (NLP) pipeline of the Engagement Analyzer consists of three components that are connected back-to-back: Transformer encoder, Span suggester, and Span categorizer. Figure 3.9 illustrates the input and output of each component, leading to the desired span annotation. In brief, the Transformer encoder takes a text and converts it into numerical values (e.g., vector representation of running tokens), the Span suggester identifies possible span candidates, and the Span categorizer predicts the probabilities of each label for each suggested candidate. The architecture of each component is described below.

Figure 3.9
NLP Pipeline component of the Engagement Analyzer.



3.3.2.3 Token embedder—Transformer encoder

The Token embedder layer takes raw textual input and produces a series of contextually aware vector representations for each token in the input text, which are used as linguistic features to predict labels in the subsequent ML pipeline. In this study, this layer is also called the Transformer encoder, because I used a family of encoder-focused masked language models (e.g., Devlin et al., 2019) for token embedding. A masked language model (MLM) is a type of pre-training language objective in NLP, which is understood as a task completed on teaching statistical patterns of sequential data (e.g., language, DNA sequences) by having them predict randomly masked tokens given contextual cues. Intuitively, MLM learns linguistic patterning by solving a vast number of cloze (or gap-filling) tasks. Variants of the MLM approach as well as their pre-trained implementations, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), are available to the research community through the HuggingFace library (Wolf et al., 2020).

Among the available MLM to date, I selected variants of A Robustly Optimized BERT Pretraining Approach (RoBERTa; Liu et al., 2019) for the current version of the Engagement Analyzer pipeline. RoBERTa is a frequently used architecture for MLM, including the out-of-the-box spaCy transformer model (i.e., en_core_web_trf). It shares most of its underpinnings with BERT but is updated in several important ways. First, although BERT was trained on a huge corpus of English texts (16GB of texts from Book Corpus and Wikipedia articles), RoBERTa was trained on an even larger corpus of English texts (over 160GB), importantly with an increased variety of text genres—Book Corpus, English Wikipedia, CC-NEWS, Open Web Text, and Stories. Liu et al. (2019) show that the increased corpus size improved the model's performance on downstream tasks such as answering questions (SQuAD; Rajpurkar et al., 2016, 2018), multi-genre natural language inferencing (MNLI-m; Williams et al., 2018), and sentiment classification (Stanford Sentiment Treebank, or SST; Socher et al., 2013). Second, RoBERTa uses a *dynamic* masking approach during training instead of the *static* masking utilized in BERT. This means that the masked tokens for cloze tasks are randomly shuffled in each training batch. Liu et al. (2019) found that *dynamic* masking improved the scores slightly compared to the static baseline. These two main updates of BERT motivated the selection of the RoBERTa architecture in the current study. As will be explained shortly, I used both the vanilla RoBERTa model available through the HuggingFace library (Wolf et al., 2020) as well as adapting this vanilla RoBERTa model to in-domain texts through a series of adaptive-pretraining steps (Ramponi & Plank, 2020). For details of this adaptive pre-training, see section 3.3.3.3.

To summarize, a variant of the RoBERTa architecture was used for the Transformer encoder, providing contextually aware vector representations of running tokens for later

components. During ML training, these Transformer layers were also fine-tuned to learn task-specific weights based on prediction losses in the Span Categorizer component.

3.3.2.4 Span Candidate Suggester

The Span Candidate Suggester component identifies candidate textual segments (i.e., spans) to be considered in the Span Categorizer for the classification of label categories. The spaCy package allows two types of implementations for this layer. The first is a machine learning approach to identify the probability of start and end tokens of each span (for a similar approach see Gu et al., 2022). The second variant uses a pre-trained dependency parser (e.g., en_core_web_trf) to identify syntactically identifiable spans (see Table 3.5 for an illustration of this approach). This latter approach can also be combined with n-gram span identification and specified n-lengths. This approach is greedier in that it recommends all textual segments with specified n-gram lengths and (predicted) dependency subtrees (here based on the spacy en_core_web_trf model; see Table 3.5). In the initial stage of the experiment, both components were tested. Although the final classification accuracy and F1 score of the downstream Span Categorizer were not affected too much by this choice, I opted for the second approach in this study for two main reasons. First, the ML approach for span identification was memory intensive, and it suggested huge numbers of spans that did not fit within available GPU memory. Second, a successfully trained ML Span Suggester only achieved approximately .7 on Recall, resulting in a ceiling effect in the subsequent Span Categorization component. Since Recall is the primary concern for the Span Suggester, I decided to use the greedier ngram + subtree suggester for all the experiments conducted below. The ngram + subtree suggester was able to cover

98.94% of actual spans in the EDT (development set), albeit with a huge number of false positives (e.g., it recommends 100 times more spans than true spans).

Table 3.5
Illustration of subtree suggester output.

Input text	Suggested Spans (all the subtrees)
My guess is that you would probably not believe in this approach yet.	[*My, My guess, guess, My guess is, is that you would probably not believe in this approach yet., My guess is that you would probably not believe in this approach yet., that, you, would, probably, not, that you would probably not believe, believe in this approach yet, that you would probably not believe in this approach yet, in, in this approach, this, this approach, approach, yet,*]

3.3.2.5 Span Categorizer

The suggested spans were considered for classification into labels in the Span Categorizer component. This final component takes a concatenated embedding for each suggested span (thus varying lengths) from the Transformer layer and applies several pooling operations and non-linear activation functions (e.g., maxout; Goodfellow et al., 2013). This layer is then sent to the densely connected linear layer, followed by the output logistic layer for final classification. The default pooling layer from the spaCy developers (Honnibal et al., 2020) is a concatenation of the following four pooling operations: the first token of the span, the last token of the span, the mean of the span, and the max of the span. This pooling layer is followed by single-layer maxout activation (Goodfellow et al., 2013) in the default setting. Finally, a logistic layer predicts the probabilities of the candidate span belonging to each category. Threshold values as well as the maximum number of predictions for each span can be set as hyperparameters for this layer (among others). One can also modify the pooling layer and non-linear activation functions. Note that, for the final output layer, the default architecture uses a series of binary classifications

instead of Softmax regression. Brief experimentation confirmed that the logistic layer worked better than Softmax layer with the range of hyperparameters used for the experiment. For this reason, I did not consider Softmax layer as an alternative architecture in the current study. Table 3.6 summarizes the major hyperparameters to tune based on the explained Engagement Analyzer component.

Table 3.6
Major hyperparameters and example values in the Engagement Analyzer pipeline.

Component	Hyperparameter	Example values
Token embedder	Pretrained model choice	roberta-base [default]; domain-adapted roberta
	Transformer encoder window	128 (stride = 96) 384 (stride = 288); 196 (stride = 128);
Span Suggester	Model architecture choice	ML architecture; ngram + dependency subtree
	Maximum n-gram lengths (when n-gram is selected)	10, 12
Span Categorizer	Span pooling layer	concat(max, mean, first, last); concat(first, last)
	Activation function	Maxout (Goodfellow et al., 2013); Mish (Misra, 2020)
	Hidden layer unit size	64 [default], 128, 256, 384, 512, etc.
	Hidden layer depth	1 [default], 2, 3, etc.
	Hidden layer dropout rate	0 [default], 0.2, etc.
	Max. number of suggestions	1, 2, etc.
	Threshold for suggestion	0.5 [default]

3.3.3 Experimental Setup

This section describes the experimental setup to train a series of machine learning models (basic architecture described in section 3.2.2) in the Engagement Discourse Treebank (EDT; see section 3.2.1). Below, I explain how I prepared the corpus splits (section 3.2.3.1), specific model architectures (section 3.2.3.2) and hyperparameters I tested (section 3.2.3.3), methods of adaptive pretraining (section 3.2.3.4), and procedures to check the stability of model performance (section 3.2.3.5).

3.3.3.1 Corpus preprocessing and splits

The Engagement Discourse Treebank (EDT) was used in the current experiment. The version of EDT used for the experiment consisted of 126,411 tokens with 4,688 sentences from the human-annotated sample of English texts in the academic domain or closely related argumentative genres (see the definition of in-domain text in section 3.2.1.1). For the current experiment, I used an 80/10/10 split to obtain a reasonable number of minority cases in the development and test set. Table 3.7 summarizes the counts of unique spans for each category. As can be seen from Table 3.7, the least frequent category was CONCUR, which only occurred 85 times in the entire dataset, followed by ENDORSE (90 times). Such imbalances across categories have repeatedly been observed in previous studies on Engagement (e.g., CONCUR occurred in only 0.81% of the entire dataset in Wu, 2007) and does not indicate a problem with the annotation.

Table 3.7
Numbers of unique tags in the corpus, and in the experimental dataset.

	Training	Dev	Test	Total	Percentage
DENY:	738	67	82	887	7.481%
COUNTER:	827	107	112	1,046	8.823%
CONCUR:	104	11	12	127	1.071%
PRONOUNCE:	259	31	28	318	2.682%
ENDORSE:	122	10	15	147	1.240%
ENTERTAIN:	2,280	269	288	2,837	23.929%
ATTRIBUTE:	887	105	108	1,100	9.278%
MONOGLOSS:	2,211	257	274	2,742	23.128%
CITATION:	482	68	68	618	5.213%
SOURCES:	707	69	79	855	7.212%
ENDOPHORIC:	162	26	25	213	1.797%
JUSTIFYING:	795	84	87	966	8.148%
Tag count	9,574	1,104	1,178	11,856	

Label imbalance is a frequently encountered problem in ML and NLP in particular (Aguiar et al., 2022). Although imbalances across the dataset do not undermine the quality of the corpus itself, they do potentially affect machine learning models during training. This is because models may perform better on the majority class and not as well on minority classes (Wang & Wang, 2022). To address this issue, various techniques have been proposed in the ML literature, including oversampling minority categories, under-sampling majority categories, and generating synthetic examples through data augmentation (for a review see Aguiar et al., 2022). In the current context, an oversampling approach was chosen because there is no reliable model to annotate or create synthetic examples and the dataset is too small to discard majority cases.

The nature of the span categorization task posed a specific challenge when conducting oversampling—data dependency. That is to say, none of the categories can be independently resampled without affecting the distribution of other categories due to the presence of multiple

labels in a given unit of analysis. Strategies to address this particular issue in resampling in sequence labeling tasks such as Named Entity Recognition (NER) seem to be emerging in the literature (Wang & Wang, 2022). Wang and Wang (2022), for example, argue that the issue of imbalanced labels in NER concerns a large number of non-labeled tokens (i.e., "O" in the IOB format). As an oversampling solution to this problem, they propose several algorithms that consider combinations of factors such as (a) count of entity tokens, (b) rarity of the label, and (c) density of NER within a sentence. Their results on several datasets indicated that the combination of (a) and (b) particularly worked well for three small datasets they investigated (but not for one large dataset), suggesting that taking the rarity of the label into account appears to be a robust approach to deal with the label imbalance problem with a small number of training sets.

Inspired by the Wang and Wang (2022) approach, the current study resampled the dataset to correct the imbalances between majority and minority labels. Since the oversampling approach in Wang and Wang (2022) is optimized with respect to empty tokens, their approach did not necessarily correct the ratios between majority and minority labels (the ratio between ENTERTAIN:CONCUR remained at 100:1.3 even after applying their method to the dataset). Thus, instead of optimizing the dataset for the number of empty tokens, I focused on ratios between majority and minority tags. This entailed the resampling of majority labels such as MONOGLOSS being conducted via minority cases only, thus narrowing the gaps in tag counts between categories.

3.3.3.2 Model Architectures

As described in section 3.2.2, the basic neural architecture of the Engagement Analyzer consists of Token Embedder, Span Suggester, and Span Categorizer. To develop a robust

machine learning model that can conduct Engagement analysis, I tested three groups of neural architectures—(a) single-transformer, (b) single-transformer+Bi-LSTM, and (c) dual-transformer+Bi-LSTM. These three groups of architecture are sketched out in Figure 3.10. For all models, spaCy's out-of-the-box model (i.e., en_core_web_trf) was used to create a dependency representation of the input sequence, which was then used to recommend span candidates (see section 3.2.2.4).

3.3.3.2.1 Single-Transformer (spaCy's default architecture)

The first group of ML models uses a single transformer layer as Token Embedder, which is then sent to a pooling layer and for logistic regression (see diagram [a] in Fig. 3.10). This is the default span categorizer implementation provided by spaCy (Honnibal et al., 2020). In my implementation, I used the en_core_web_trf model to predict dependency representations of the input text, which were used to suggest candidate spans. Each candidate span's representation is then created by taking the RoBERTa embedding and applying several pooling operations. The pooled span representation is sent to the non-linear activation function and subsequently to the logistic layer for prediction. In this architecture, the RoBERTa embeddings were fine-tuned to learn task-specific weights while the Transformer layer from en_core_web_trf was fixed.

3.3.3.2.2 Single-Transformer + Bi-LSTM

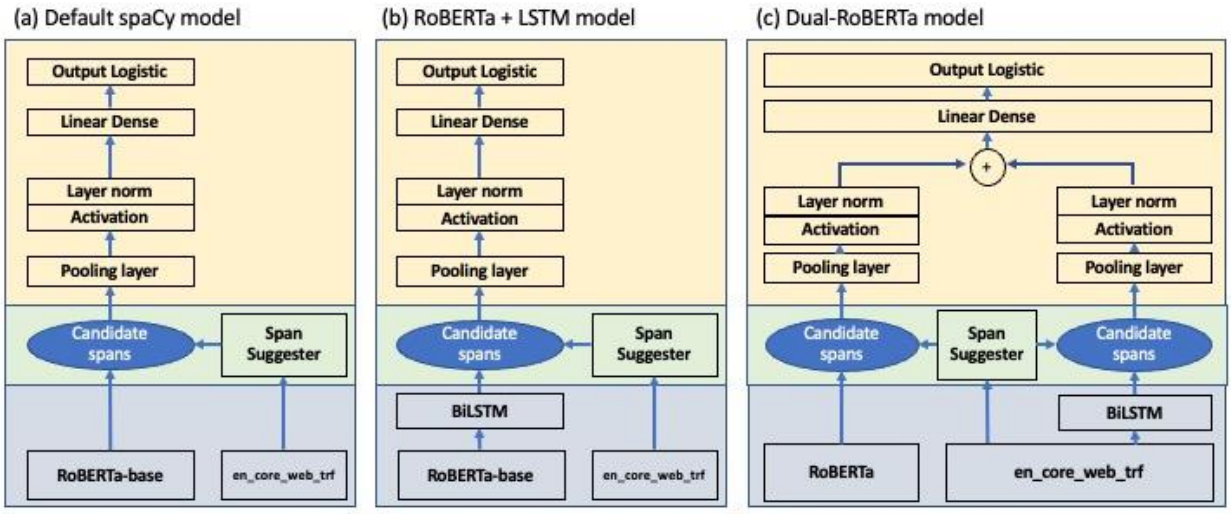
Although transformer embedding can provide a contextually aware token representation (Clark et al., 2019), it was hypothesized that additional sequential information might be beneficial for the classification of a specific category of engagement, such as ATTRIBUTION. To allow the model to learn this additional contextual information, I added a single-layer Bidirectional Long-Short Term Memory (Bi-LSTM; Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) architecture on top of the RoBERTa embeddings, before they were sent to the

span pooling layer. This architecture has often been implemented in previous NLP tasks involving span identification (e.g., Gu et al., 2022; K. Lee et al., 2017; Papay et al., 2020; Zhu et al., 2021). For the current study, I used one-layer Bi-LSTM with 200 hidden dimensions following a previous study (Gu et al., 2022).

3.3.3.2.3 Dual-Transformer + single-layer Bi-LSTM

The third architecture used two sets of transformer embeddings side-by-side, concatenated before the final output layer for prediction (see Architecture (c) in Fig. 3.10). This model architecture was inspired by recent ensemble approaches to span identification pipelines (e.g., Rao, 2022). The intuition underlying the dual-Transformer architecture was that the two Transformer models would offer complementary information to categorize the span labels. In the current implementation, this architecture is an extension of (a) the single-transformer model. The first path (in the left of the panel (c) Fig. 3.10) mirrors model (a) in that the vanilla or domain-adapted RoBERTa models provide token embeddings of the input sequence, and their weights are fine-tuned. The second transformer model (in the right of the panel (c) Fig. 3.10), which was used solely for span suggestion by creating dependency representation (i.e., en_core_web_trf), was then used as an additional feature extractor. Arguably, the en_core_web_trf model may provide additional information because its RoBERTa embedding was already fine-tuned under the multi-task learning setting in the OntoNote corpus for POS tagging, dependency parsing, and Named Entity Recognition. Thus, it was hypothesized that the en_core_web_trf model would provide additional information above and beyond pre-trained MLM, thus boosting the accuracy of identifying spans that tend to be associated with distinct syntactic structures (e.g., JUSTIFYING).

Figure 3.10
Three alternative neural architectures for the proposed Engagement Analyzer.



3.3.3.3 Hyperparameters

As listed in Table 3.6, major hyperparameters to tune included the choice of pre-trained Transformer encoders, pooling operations, the nonlinear activation function, their hidden sizes, and their dropout rate. Table 3.8 lists the ranges of hyperparameters on which a search was conducted. I conducted a hyperparameter search on each of the three architectures with the same range of parameter values. Note that the Dual-Transformer model has two separate Feed-Forward Networks (FFNs). All runs used the Adam optimizer with beta1 = 0.9, beta2 = 0.999, epsilon = 1e-8, and L2 regularization with 0.01 (treated as a weight decay term). The learning rate was scheduled as linearly decaying with a max step of 20,000 after a linear warm-up using the first 1,000 training steps. As shown in Table 3.8, the maximum learning rate was randomly chosen under a uniform distribution ranging from 2e-5 to 6e-5. As a side note, the spaCy package allows sampling mini-batches for training with a specified number of words instead of documents. In this experiment, word-based batching strategies were used.

For the selection of the Transformer encoder, I tested three variants—RoBERTa-base, RoBERTa-student-writings (SW), and RoBERTa-academic (A). The first, RoBERTa-base (Liu et al., 2019), is a vanilla RoBERTa distributed through HuggingFace (Wolf et al., 2020). Although the vanilla RoBERTa-base model is known to perform well on a variety of NLP tasks since its introduction, one potential limitation of this for the current study is that it was not trained on the in-domain text of the current study. This may result in sub-optimal performance of the final ML model due to this domain mismatch. To overcome this potential problem, I prepared a domain-adapted version of the RoBERTa model by resuming the training on academic texts (see section 3.2.3.4 for a domain adaptation experiment).

I used three versions of the non-linear activation layer. First, I used the maxout activation layer (set by spaCy) (Goodfellow et al., 2013). The second set of the activation function was the Mish activation function (Misra, 2020). This was considered here because it outperformed the maxout activation function, particularly for minority categories in a preliminary experiment. I also tested the Gelu function in this preliminary experiment; however, it did not improve the model's performance. For the two activation functions, I modified the original spaCy code to allow different dropout rates as well as the depth of the network (with layer normalization for each).

Finally, I implemented a variant of Mish activation function layers where two independent hidden layers were concatenated. The last one was mainly motivated by an observation in the early stage of the experiment that the model tended to have trade-offs between different categories. It was hypothesized that separate layers with slightly different input information would improve a specific part of the model (allowing for each path to specialize in various aspects of the embedding, for example). Thus, for the pooling layer, I prepared two

different implementations. The first one—a concatenation of the first token, the last token, the mean of the entire span, and the max of the whole spans—is a default spaCy implementation, and it is used as the input for all three versions of the activation function outlined. Additionally, I prepared a slightly modified version of the two-way activation layer for one of the activation paths, where I used the sum of the span instead of the max. The second version of the pooling operation should have more information about the entire span rather than specific parts of it, potentially impacting on the behavior of longer sequences.

Table 3.8

List of hyperparameters used to train the end-to-end Engagement Analyzer pipelines.

Category	Hyperparameter	Range or Choice		Selection
Entire model	Model Architecture	Single-Transformer; Single-Transformer+ LSTM; Dual-Transformer + single-LSTM		discrete
Token Embedder	Pre-Trained language model	roberta-base; egumasa/roberta-base-academic3; egumasa/roberta-base-university-writing2; egumasa/roberta-base-research-papers		discrete
Span Categorizer	FFN (Activation function)	Maxout; Mish; Mish with two separate FFNs		discrete
Span Categorizer	FFN (hidden unit sizes)	[128, 256, 384]		discrete
Span Categorizer	FFN (dropout rates)	[0, 0.2, 0.3, 0.4]		discrete
Span Categorizer	FFN (layer depths)	[1, 2]		discrete
Training	Maximum learning rate (alpha)	6e-5 – 2e-5		uniform distribution
Training	System seed during training	[0, 808, 1993, 1234, 2023]		discrete
Training	Gradient accumulation steps	[4, 8]		discrete
Span Suggester	Max n-gram lengths		12	fixed
Training	Optimizer	Adam with weight decay		fixed
Training	Learning rate schedule	linear decay with warm-up steps		fixed
Training	Warm-up steps		1,000	fixed
Training	Maximum training steps		20,000	fixed
Training	Steps before early stop		3,000	fixed
Training	mini-batch size	defined by number of words		fixed
Training	minimal start batch size	[300, 500, 900]		discrete
Training	Maximum batch size	1,000 words		fixed

Note. Category indicates the facets of the model for which the hyperparameter is relevant; FFN refers to Feed-Forward Network; Range or Choice column lists possible hyperparameter choices; Selection column indicates whether the parameter is fixed across models, selected randomly from a discrete choice, or selected randomly from a specified distribution.

The first domain-adapted version—RoBERTa-student-writing—was refined on the same five corpora used to sample the EDT (i.e., BAWE, MICUSP, TOEFL11, CLC FCE, and ICNALE; see section 3.2.1.2 for descriptions). RoBERTa-student-writing was trained on a directly relevant dataset for the fine-tuning objective. The third variant, RoBERTa-academic, is another domain-adapted version trained on a combination of the five corpora sampled for EDT and the Elsevier OA CC-BY Corpus (Kershaw & Koeling, 2020). The Elsevier OA CC-BY Corpus is a collection of 40,001 open-access (OA) articles from Elsevier journals across various disciplines. It was added to this experiment to represent the professional discourse that student writers (one of the stakeholders of the Engagement Analyzer) would socialize with. In adapting the model, the original RoBERTa checkpoint was accessed through the HuggingFace library. For the RoBERTa-Academic model, I continued training the model for an additional 20 epochs, with effective batch sizes of 512 (8 batches with 64 gradient accumulation steps). I used the Adam optimizer with beta1 = 0.9, beta2 = 0.99, and epsilon = 1e-08. For the learning rate schedule, the maximum learning rate was set to 7e-5, with linear decay and 20% of the steps (i.e., 4 epochs) were used for the warm-up. For the RoBERTa-student-writing model, I trained the original RoBERTa-base for an additional 20 epochs, with an effective batch size of 128 (8 with 16 gradient accumulation steps). I use Adam with the same setting, with a learning rate of 5e-5 linearly decaying over time with 10% of the entire training steps used for the warm-up.

To show whether the aforementioned domain adaptation resulted in better representation of academic discourse features (i.e., word choice), Table 3.9 lists the top 10 candidate tokens returned by each of the three models in the mask-filling task on the following stem—"The goal of this paper is to <mask>". As predicted, the off-the-shelf RoBERTa-base model tended to recommend words that are characteristic of general writing (e.g., *help*, *inspire*, *inform*), which

may not necessarily reflect the research activity itself. On the other hand, both the domain-adapted versions of RoBERTa appear to show more awareness of academic word choice—*explain*, *discuss*, and *understand* as the top 3 candidates in the RoBERTa-SW model, and *describe*, *explain*, and *discuss* in the RoBERTa-A model. Further, while the RoBERTa-SW model tended to give slightly higher weights to "explain", the RoBERTa-A model exhibited more even probability distributions of alternative verbs that fill the mask, implying a better representation of a range of academic discourse features.

Table 3.9
Top 10 predicted tokens for each of the RoBERTa model variants on the stem "The goal of this paper is to <mask>".

	RoBERTa-base		RoBERTa-student-writing (SW)		RoBERTa-Academic	
Rank	Predicted token	Prob	Predicted token	Prob	Predicted token	Prob
1	help	0.300	explain	0.119	describe	0.093
2	clarify	0.101	discuss	0.095	explain	0.089
3	explain	0.044	understand	0.082	discuss	0.073
4	improve	0.044	clarify	0.077	clarify	0.037
5	inspire	0.044	help	0.052	illustrate	0.034
6	educate	0.035	illustrate	0.050	investigate	0.033
7	inform	0.033	investigate	0.031	present	0.029
8	communicate	0.022	summarize	0.031	understand	0.029
9	summarize	0.021	improve	0.028	identify	0.028
10	understand	0.019	describe	0.027	summarize	0.026

3.3.3.5 Hyperparameter search with random restart and 5-fold Cross-validation

To answer RQ2—What are the impacts of machine learning architecture selection and associated hyperparameters on precision, recall, and F1 scores?—I conducted a comprehensive hyperparameter search to determine whether any of the architectures outperformed others. Specifically, given the hyperparameter I presented in section 3.2.3.3, I conducted a random search with random restarts and five different seeds. Random search, where a combination of

hyperparameters is randomly chosen for each model run, enables a comprehensive search. I also used five seeds to randomize the dataset during training, allowing some variability in each run of the training sample. The result of this hyperparameter search allowed statistical tests of hyperparameter settings. Thus, the best model was chosen by conducting mixed-effect modeling on the scores.

Given the small size of the training dataset, it is possible that the development dataset does not capture the entire data distribution, resulting in biases in model accuracy. To counteract this problem, I further checked the stability of the candidate model based on the hyperparameter search through five-fold CVs. To this end, I divided the data into roughly five equal sets, I trained five equivalent models using different held-out data for development and testing each time. Subsequently, the median, as well as the variability in scores, was inspected for potential biases in the candidate hyperparameter settings. For those settings that produced acceptable ranges of performance, I treated the original model trained on the 80/10/10 split as the final production model.

3.3.3.6 Evaluation metrics

In the current study, two sets of evaluation metrics were used—Macro averages of precision, recall, and F1 and Cohen's kappa. Precision, Recall, and F1 scores are commonly used evaluation metrics in the machine learning literature. Precision and Recall are calculated for each category using the following formula:

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

where True Positives (TP) are defined as the number of candidate spans (not tokens) correctly classified for a given category; False Positives (FP) concern the number of candidate spans incorrectly classified in a given category; and False Negatives (FN) count the number of candidate spans incorrectly labeled as other categories (including the empty label "O"). Accordingly, the fewer FPs identified by the system, the greater will be the precision. Similarly, the fewer the true spans missed by the system, the higher will be the recall score.

F1 score is a harmonic mean of recall and precision calculated using the following formula:

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Thus, the F1 score is considered to be a balanced single metric to summarize Precision and Recall scores and is often introduced as a metric to evaluate overall system performance in the machine learning literature (see Dror et al., 2020; Thampi, 2022).

Since Precision, Recall, and F1 scores are calculated per label, it is customary to summarize the quantities into a single metric by averaging over the categories. Two ways to average by-category scores are often used for different purposes—Macro and Micro averages. Macro averages are a simple average over the number of categories as the denominator, having equal weights for each category in the calculation. This may be desirable when one cares about the performance of minority labels. Micro averages are calculated at the item level regardless of the category boundary, reflecting the proportions of categories propagated to average calculation. Micro averages should be a more direct reflection of the probability of a given label being correct regardless of the categories being considered. In this study, I mainly focus on macro averages so as to give equal consideration to majority and minority categories in the evaluation.

Although precision, recall, and F1 scores are widely used as standard metrics to evaluate a model's performance, they do not consider chance-level agreement between gold-tags and system predictions. For this reason, I also used Cohen's kappa coefficient (Cohen, 1960) to supplement interpretations of model performance. As explained by McHugh (2012) and Landis and Koch (1977), Cohen's kappa coefficient (Cohen, 1960) ranges from -1 to +1, and is closely related to the intra-class correlation coefficient. It adjusts observed agreement rates based on expected agreement rates for a given contingency table, as in the following formula:

$$Kappa = \frac{\Pr(actual) - \Pr(expected)}{1 - \Pr(expected)}$$

Where the expected agreement rate is calculated using the total counts of independent observations and marginal sums of the contingency table. Taking a simple 2-by-2 contingency table as a hypothetical example, expected agreement is calculated as (after McHugh, 2012):

$$\Pr(expected) = \frac{\left(\frac{Column\ 1\ marginal * Row\ 1\ marginal}{n}\right) + \left(\frac{Column2\ margninal * Row2\ marginal}{n}\right)}{n}$$

Suppose that the hypothetical chance level agreement is .20 and the observed agreement rate is .74, then the kappa coefficient would be:

$$kappa = \frac{.74 - .20}{1 - .20} = .675$$

Resulting in an adjusted rate of agreement given the chance level agreement.

Table 3.10 summarizes two existing benchmarks to interpret the level of agreement based on Cohen's kappa values. In Landis and Koch's (1977) classic benchmark, a kappa coefficient of above .60 is considered substantial. In contrast, McHugh (2012) proposes a more conservative benchmark, treating the .60-.79 range as moderate agreement, and .80-.89 as strong agreement.

Following these two guidelines, I consider a kappa value below .6 to be "inadequate" (McHugh, 2012, p. 279), kappa between .60 and .80 as "moderate-to-substantial", while a value over .80 indicates "strong" reliability of an instrument.

Table 3.10
Two existing benchmarks for interpreting Cohen's Kappa coefficient.

Kappa values	Landis & Koch's (1977) benchmark	McHugh's (2012) revised benchmark
< .00	Poor	None
.00 – .20	Slight	None
.21 – .39	Fair	Minimal
.40 – .59	Moderate	Weak
.60 – .79	Substantial	Moderate
.80 – .89	Almost Perfect	Strong
> .9	Almost Perfect	Almost Perfect

3.4 Results

3.4.1 RQ1: What is the intercoder agreement rate for Engagement annotation?

The first research question concerned the intercoder agreement between two trained annotators. As described in section 3.3.1.3, the two annotators were undergraduate students majoring in linguistics, and they underwent an intensive training procedure that lasted for more than one term (see Figure 3.8). Intercoder agreement thus can be considered as the level of agreement between human coders with some background in linguistics and sufficient practice.

As shown in Table 3.11, the overall Cohen's kappa coefficient was .6708 (substantial; Landis & Koch, 1977; moderate; McHugh, 2012), and macro F1 was .67. Table 3.11 also presents precision, recall, and F1 scores using Annotator B as the reference. Several patterns emerged from the intercoder agreement. First, the agreement was strong for six categories (F1 > .80). Second, the coders tended to disagree on the four remaining categories. These were ATTRIBUTION (F1 = .60), ENDOPHORIC (F1 = .62), PROCLAIM, (F1 = .40), and

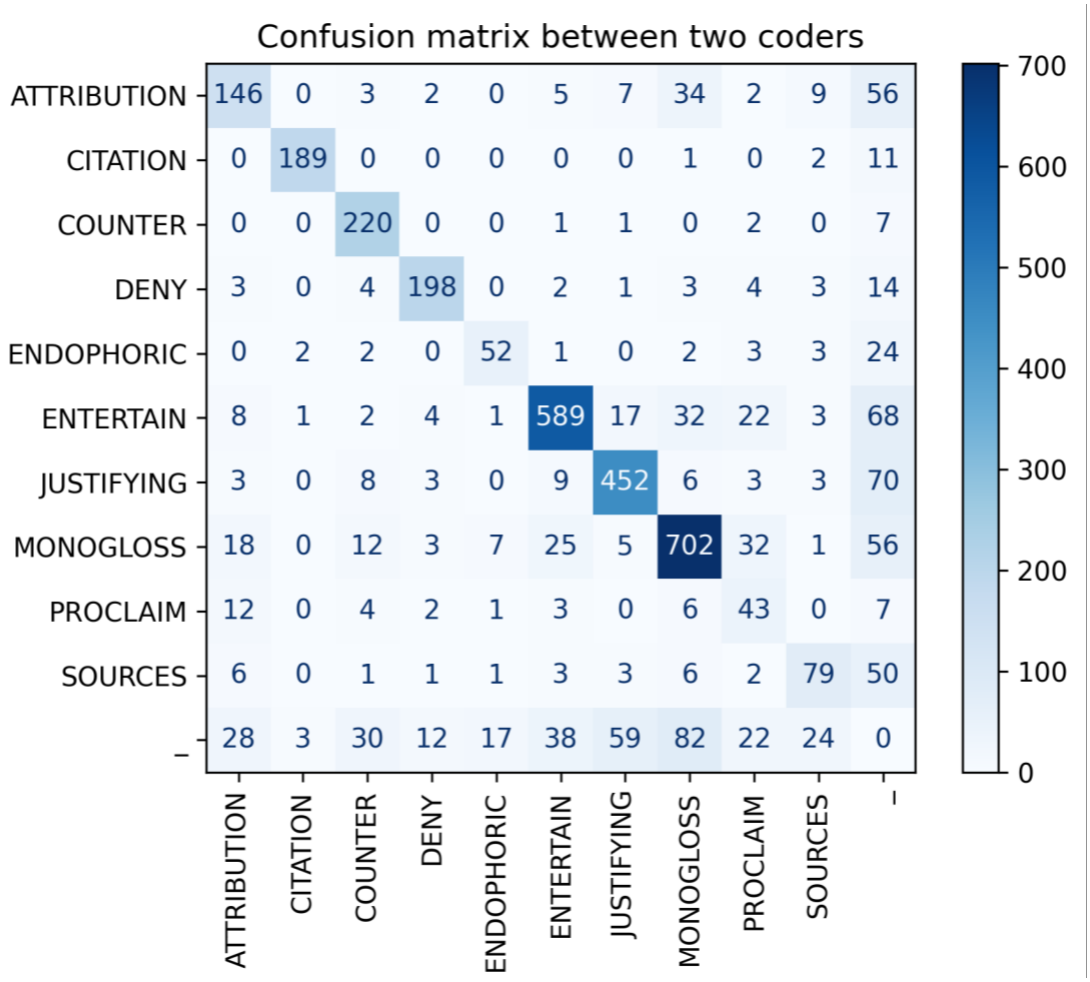
SOURCES (F1 = .57). Looking at the confusion matrix (Figure 3.11), there seemed to be several categories that were particularly difficult to distinguish from one another—ATTRIBUTION and MONOGLOSS, ATTRIBUTION and PROCLAIM, ENTERTAIN and MONOGLOSS. In other cases, difficulties mostly stemmed from the detection of a rhetorical move itself (e.g., ENDOPHORIC vs no tags).

Given the moderate-to-substantial intercoder agreement (Cohen's kappa = .6708; F1 = .67), I suggest that annotating Engagement resources is a relatively difficult, albeit not impossible, task to do, given sufficient time to train the coders (approximately 10–20 weeks of hands-on training). The above findings also provide a reasonable benchmark score against which machine-learning models should be compared. I suggest that machine learning models should perform similarly to trained human coders if they score above kappa of .67 and macro F1 = .67. Based on intercoder agreement, I also expect that ML models would struggle to identify categories such as ATTRIBUTION, ENDOPHORIC, PROCLAIM, and SOURCES.

Table 3.11
Intercoder agreement (using Annotator B as a reference).

	F1 scores reported in Read and Carroll (2012)	precision	recall	f1-score	support
ATTRIBUTION	.427	0.65	0.55	0.6	264
CITATION	*NA*	0.97	0.93	0.95	203
COUNTER	.603	0.77	0.95	0.85	231
DENY	.451	0.88	0.86	0.87	232
ENDOPHORIC	*NA*	0.66	0.58	0.62	89
ENTERTAIN	.459	0.87	0.79	0.83	747
JUSTIFYING	*NA*	0.83	0.81	0.82	557
MONOGLOSS	*NA*	0.8	0.82	0.81	861
PROCLAIM	.246	0.32	0.55	0.4	78
SOURCES	*NA*	0.62	0.52	0.57	152
_		0	0	0	315
accuracy				0.72	3729
Cohen's kappa				0.67	
macro Avg		0.67	0.67	0.67	3729
weighted Avg		0.73	0.72	0.72	3729

Figure 3.11
Confusion matrix for two sets of annotation.



3.4.2 RQ2: What are the effects of different architecture and hyperparameter settings on Precision, Recall, and F1 scores?

The second aim of this study was to investigate the effects of architecture and hyperparameter choices for the Engagement Analyzer. To this end, I trained three separate ML architectures with varying hyperparameter settings with the *weights and biases* (wandb) package in Python. The wandb package integrates with the spaCy package, allowing tracking of ML experiments and providing a random parameter search capability. For each of the three ML

architectures outlined in section 3.3.3.2, I ran a random search with Bayesian optimization (see Table 3.8 for these hyperparameter spaces). Using this strategy, I trained a total of 205 models, and the resulting models were evaluated against the development set. In what follows, I present the results of descriptive and inferential statistics for the following major hyperparameters—Architectures, Pre-trained RoBERTa model, and Activation functions. For each category of hyperparameters, I first examined the results for overall precision, recall, and F1 scores and then by-category scores. For visualization and simple inferential statistics, I used the *ggplot2* and *ggstatplot* packages in R. The *ggstatplot* package allows visualization and simple statistical testing simultaneously. In these simple inferential statistics, I conducted a series of Bayesian one-way analyses of variance for each major hyperparameter setting and computed Bayes Factors. A logarithm of Bayes Factors was used to interpret magnitude on a linear scale in both directions, using Raftery's (1995) benchmark: Weak ($0 < \log(BF) \leq 1.098$), Positive ($1.098 < \log(BF) \leq 2.9957$), Strong ($2.9957 < \log(BF) \leq 5.0106$), and Very strong ($\log(BF) > 5.0106$).

Because a simple test does not account for any effects of potential confounders or the by-category performance of the chosen hyperparameter, I also conducted a mixed-effects linear regression analysis if the findings held even after controlling for confounds. The following formula describes a generic model I fitted to investigate the effects of major hyperparameters:

- lmer(scores ~ metric * split * category * architectures * `Transformer` + scale(`training.optimizer.learn_rate.initial_rate`) + (1|dropout) + (1|hidden_sizes) + (1|reducer) + (1|seed), data = data_long)

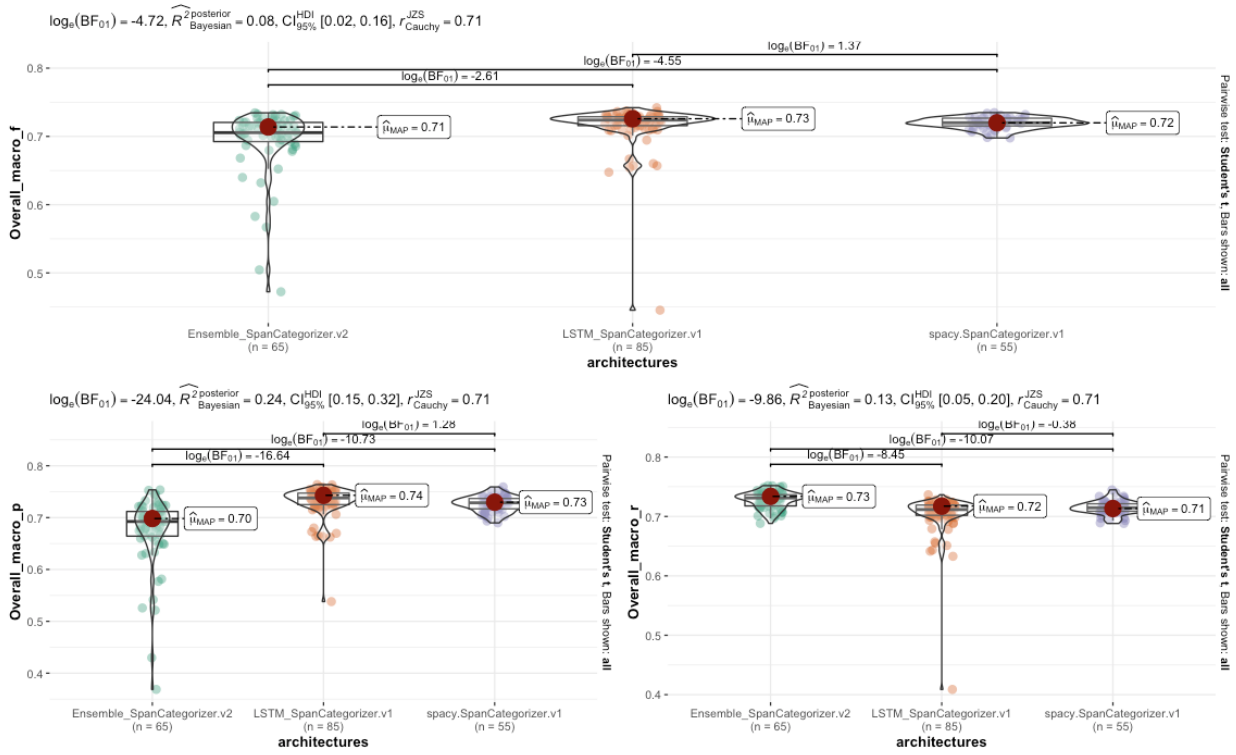
Scores were either precision, recall, or F1 scores for each category of respective runs. Metric (nominal variable) specifies whether the outcome variable is precision, recall, or F1 score, and Category specifies which Engagement categories of the outcome variable. Architecture is a

three-level factor, specifying which architecture the model was trained on. Transformer specifies the selection of pre-trained RoBERTa models. These five-way interactions allowed examining the interplay of these variables. The maximum learning rate (from 6e-5 to 2e-5; see Table 3.8) was used as a fixed effect parameter to control for the effects of the maximum learning rate of each run. Four random effects were entered to account for any other effects during model training —dropout rate, hidden unit sizes, the choice of Feed-Forward Network, and random seeds used during training. Since the primary interest was to find out any effects of chosen hyperparameters while controlling for others, I did not interpret the five-way interaction in the model input. Instead, a post hoc comparison was made and reported using the *emmeans* package.

3.4.2.1. Architecture selection

As can be seen in Figure 3.12, the selection of model architecture significantly impacted on the performance of the NLP pipeline. The result of a Bayesian one-way ANOVA (Fig. 10), as well as mixed-effects modeling, showed that the dual-transformer models underperformed the other two architectures on overall F1 (positive-to-strong evidence) and Precision (strong evidence). The dual-transformer architecture also scored higher on Recall with very strong evidence. It is also worth noting that the dual-transformer architecture showed higher variability in performance compared to the other two architectures. This implies that the selections of other hyperparameters greatly influenced the model performance of the dual-transformer model.

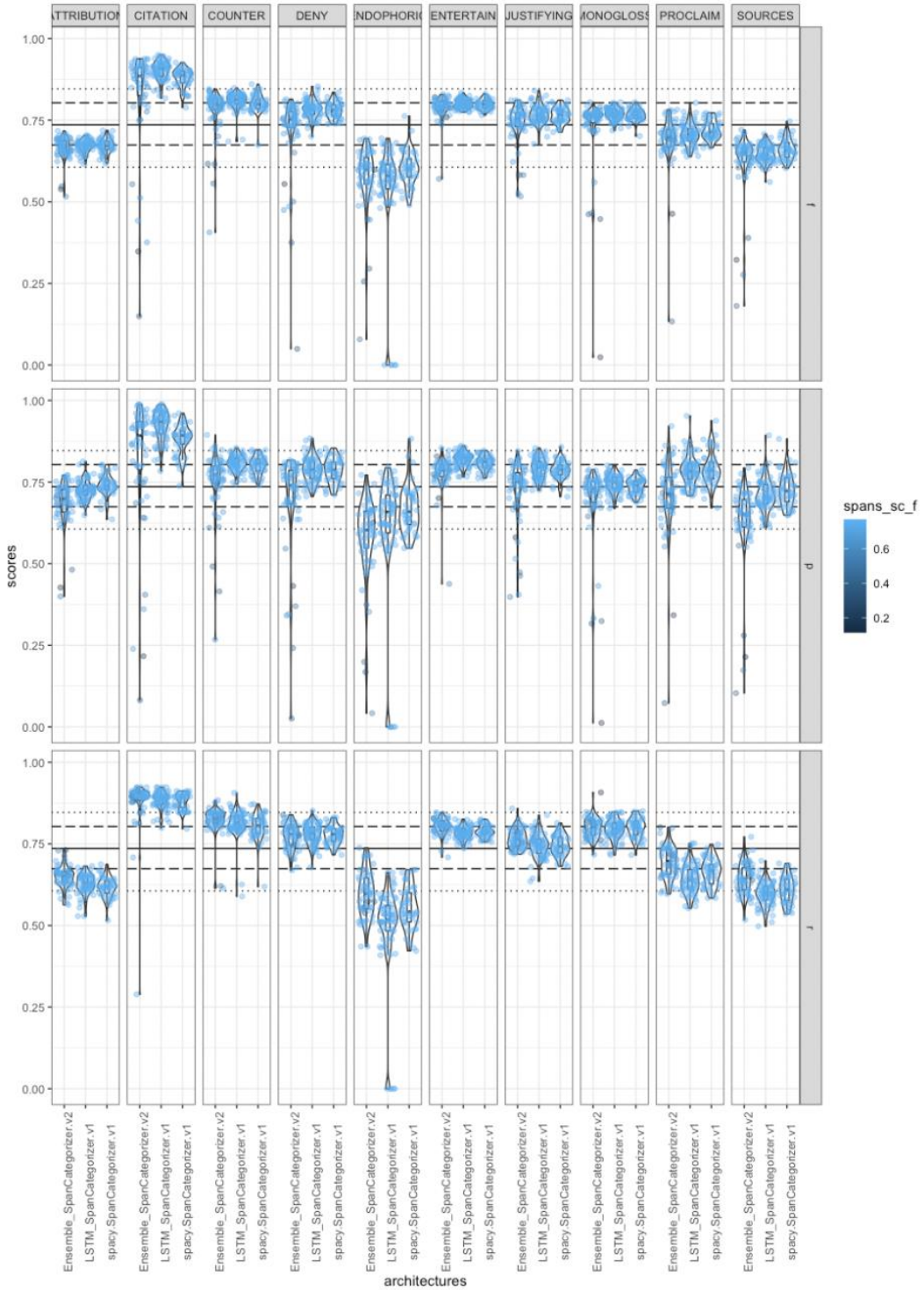
Figure 3.12
Overall performance comparisons of three architectures.



To further investigate the by-category performance of the three architectures, I visually inspected their precision, recall, and F1 scores (see Fig. 3.13). Overall, this confirmed the overall finding that the dual-transformer model showed high variability across runs. In some cases, the dual-transformer models showed the highest recall for specific categories (e.g., ATTRIBUTION, PROCLAIM, and SOURCES in the leftmost box in the lowest panels). There were a few by-category outliers (e.g., overachievers) in the single-transformer model (e.g., ENDOPHORIC and SOURCES). This general pattern was confirmed in a mixed-effects linear regression, where the dual-transformer architecture tended to underperform the other two architectures in the CITATION, DENY, JUSTIFYING, MONOGLOSS, PROCLAIM, and SOURCES tags (see Fig. 3.14 for model-based by-category comparison).

Figure 3.13

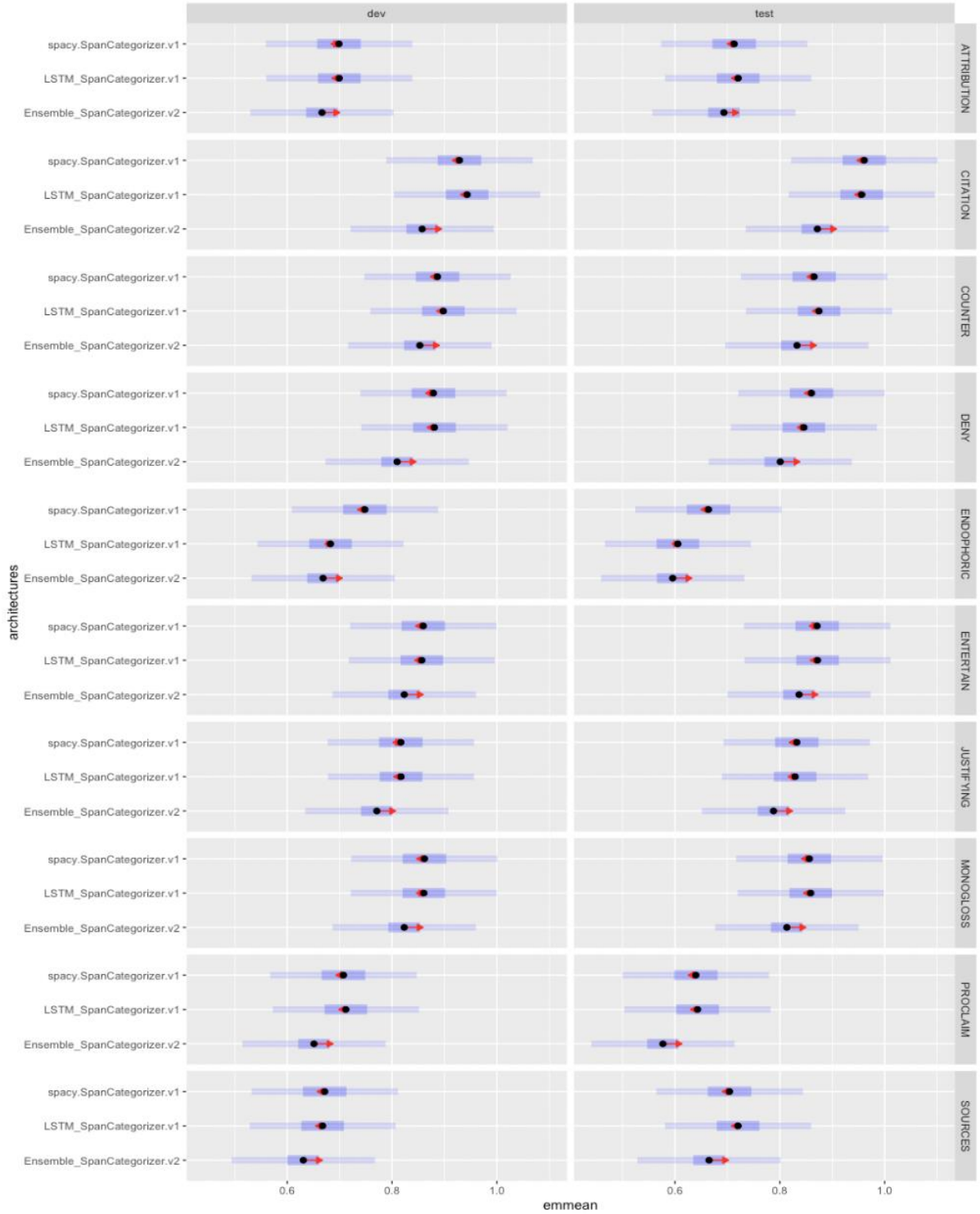
By-category performance comparison of three architectures.



Note. The darker the dots, the lower the overall F1 scores. The solid horizontal line shows the mean of the entire by-category scores. The longer dashed lines represent 27% and 75% percentiles; The dotted lines represent the 10% and 90% percentiles of by-category scores.

Figure 3.14

Model-based comparisons of the by-category performance of three architectures.



Note. The black dots show estimated marginal means; a pair of non-overlapping red allows show statistical significance; the thicker ribbons show Confidence Intervals (CIs) and the thinner ribbons represent prediction intervals (PIs).

3.4.2.2. Pre-trained RoBERTa model

As can be seen from Figure 3.15 (Bayesian one-way ANOVA) and Figure 3.16 (marginal means of the mixed-effect model), none of the comparisons resulted in strong evidence favoring a particular version of the pre-trained model.

Figure 3.15

Overall performance comparison of different pre-trained RoBERTa models.

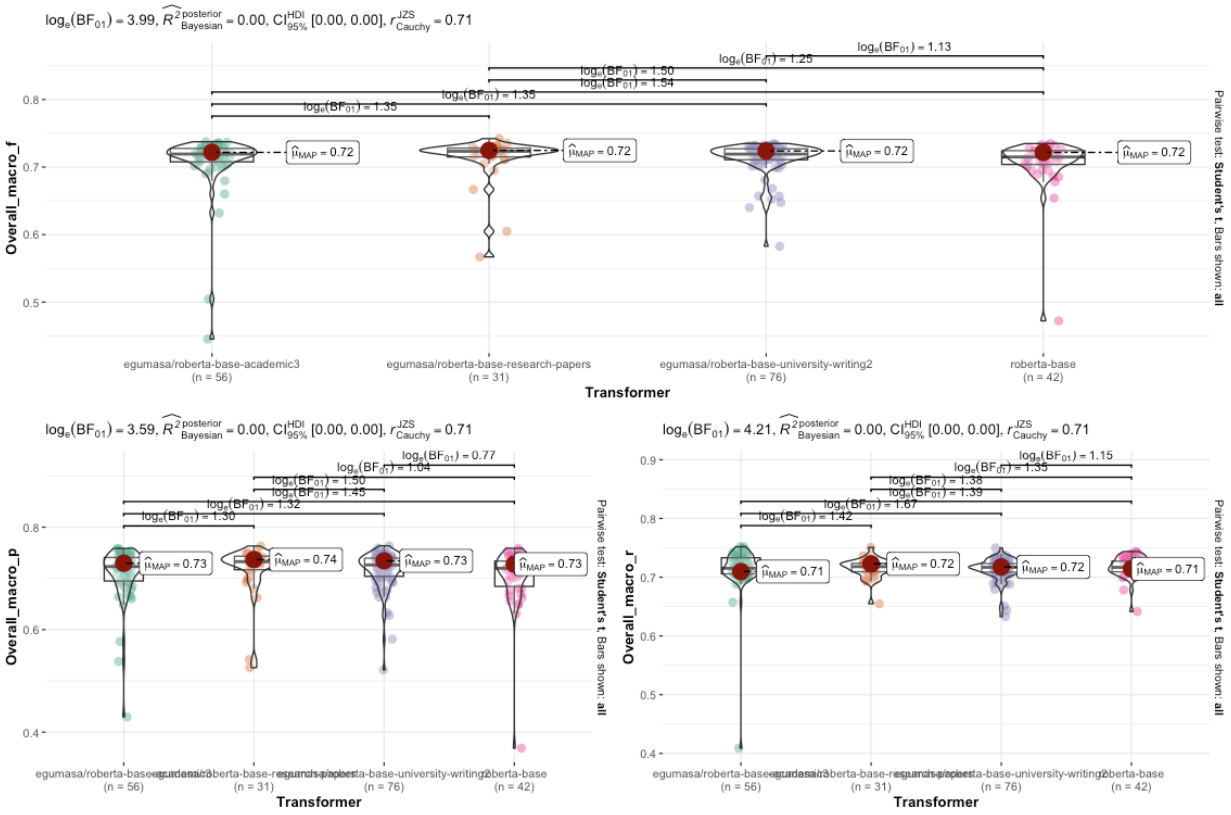
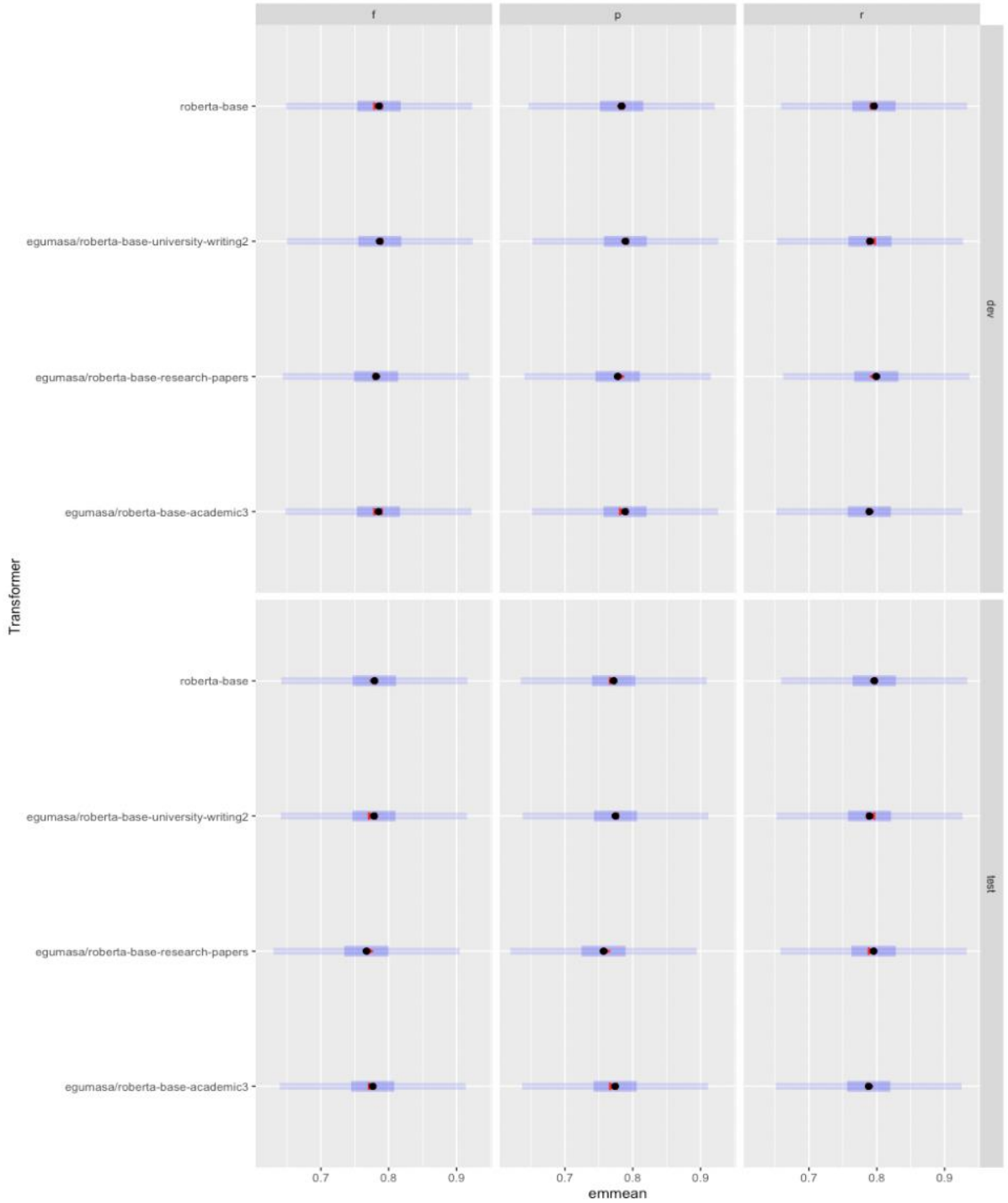


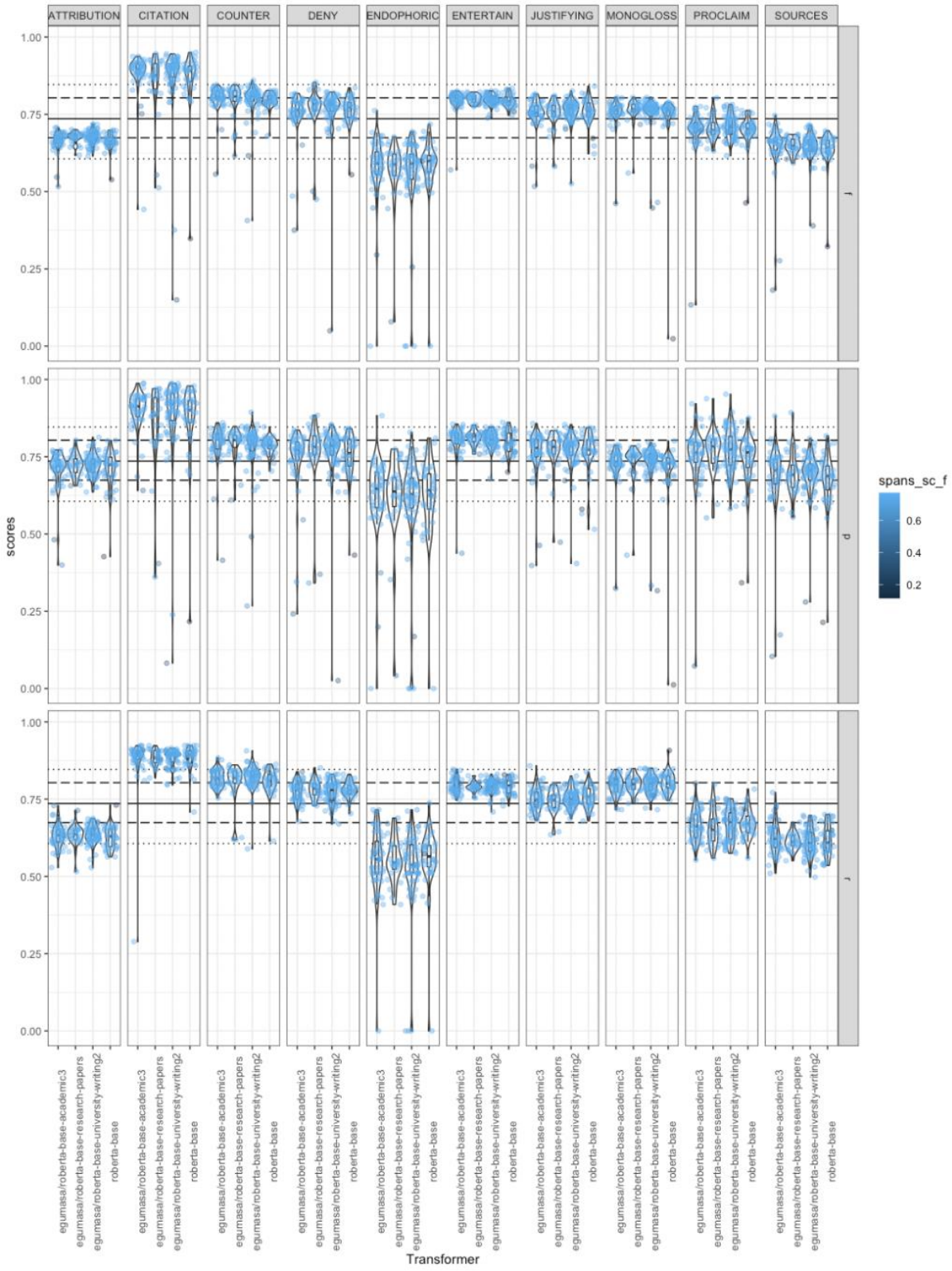
Figure 3.16

Model-based comparisons of overall performance of the four RoBERTa models chosen.



Note. The black dots show estimated marginal means; a pair of non-overlapping red allows show statistical significance; the thicker ribbons show Confidence Intervals (CIs) and the thinner ribbons represent prediction intervals (PIs).

Figure 3.17
By-category performance comparisons of four RoBERTa models.



3.4.2.3. Activation functions and their depths and hidden sizes

Next, I compared performances with different activation functions, their depths, and hidden sizes. Note that formal comparisons were only conducted on the two architectures that used a single Feed-Forward Network—i.e., the single-Transformer and Transformer+LSTM architectures. As shown in Figure 3.18, no pair was strongly different from any other (logged BF < 3). Figure 3.19 shows a comparison of hidden layer unit sizes. Logged BF showed strong evidence to favor a unit size of 384 over 128, while the other pairs were not strong.

Figure 3.18

Overall performance comparison of Activation functions for single-transformer and Transformer + LSTM models (F1 score).

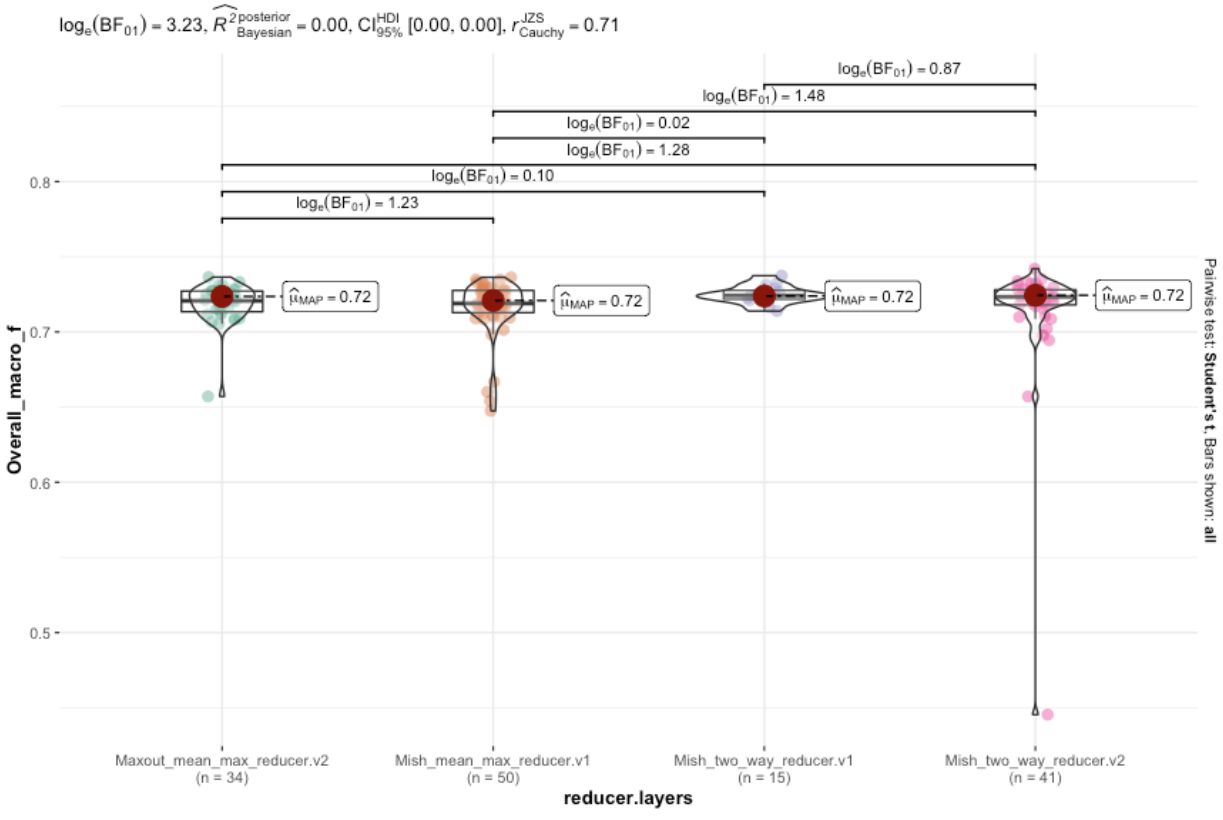
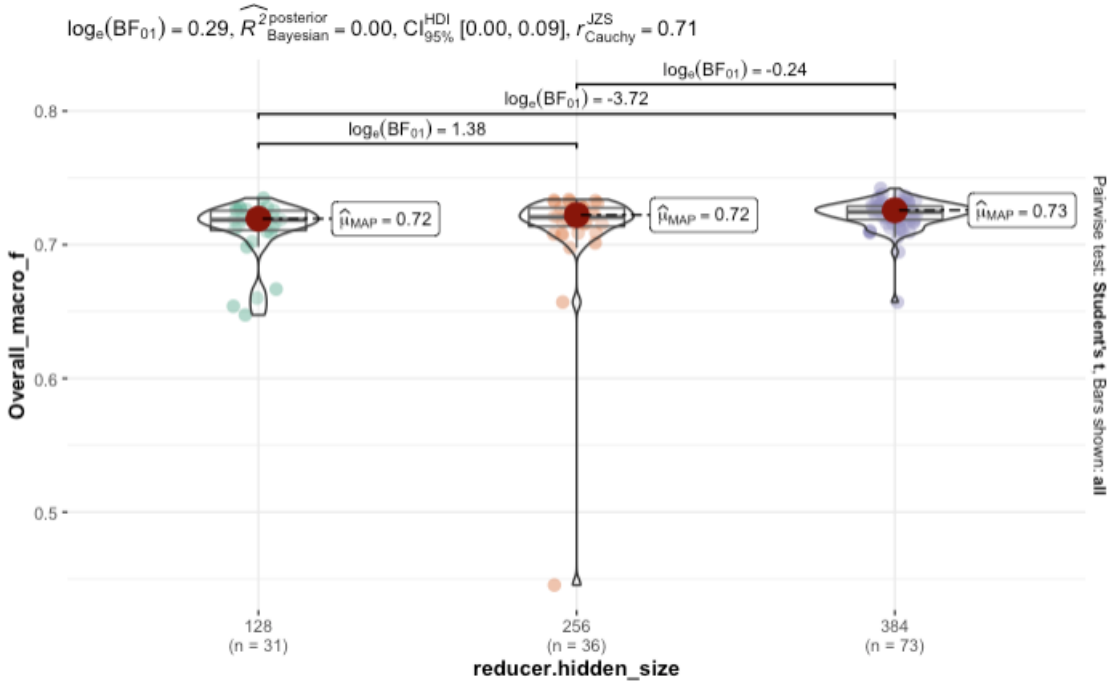


Figure 3.19

Overall performance comparison of hidden unit sizes.



3.4.2.4. Interim summary

The results for RQ2 indicated that, overall, the dual-Transformer model tended to underperform the other two models, particularly in the Precision metric. The dual-Transformer model, however, tended to significantly outperform the other two on the Recall metric. I did not observe a relative benefit of using a particular version of the pre-trained RoBERTa model when other variables were constant. Finally, the effects of the FFN were minimal, and there was a tendency for larger hidden sizes (i.e., 384) to outperform smaller ones (i.e., 128). This suggests that the effects of individual hyperparameters were not influential on overall performance; however, questions remain about whether there is any combination of hyperparameters that results in the best performance. This issue is the scope of the next question (RQ3).

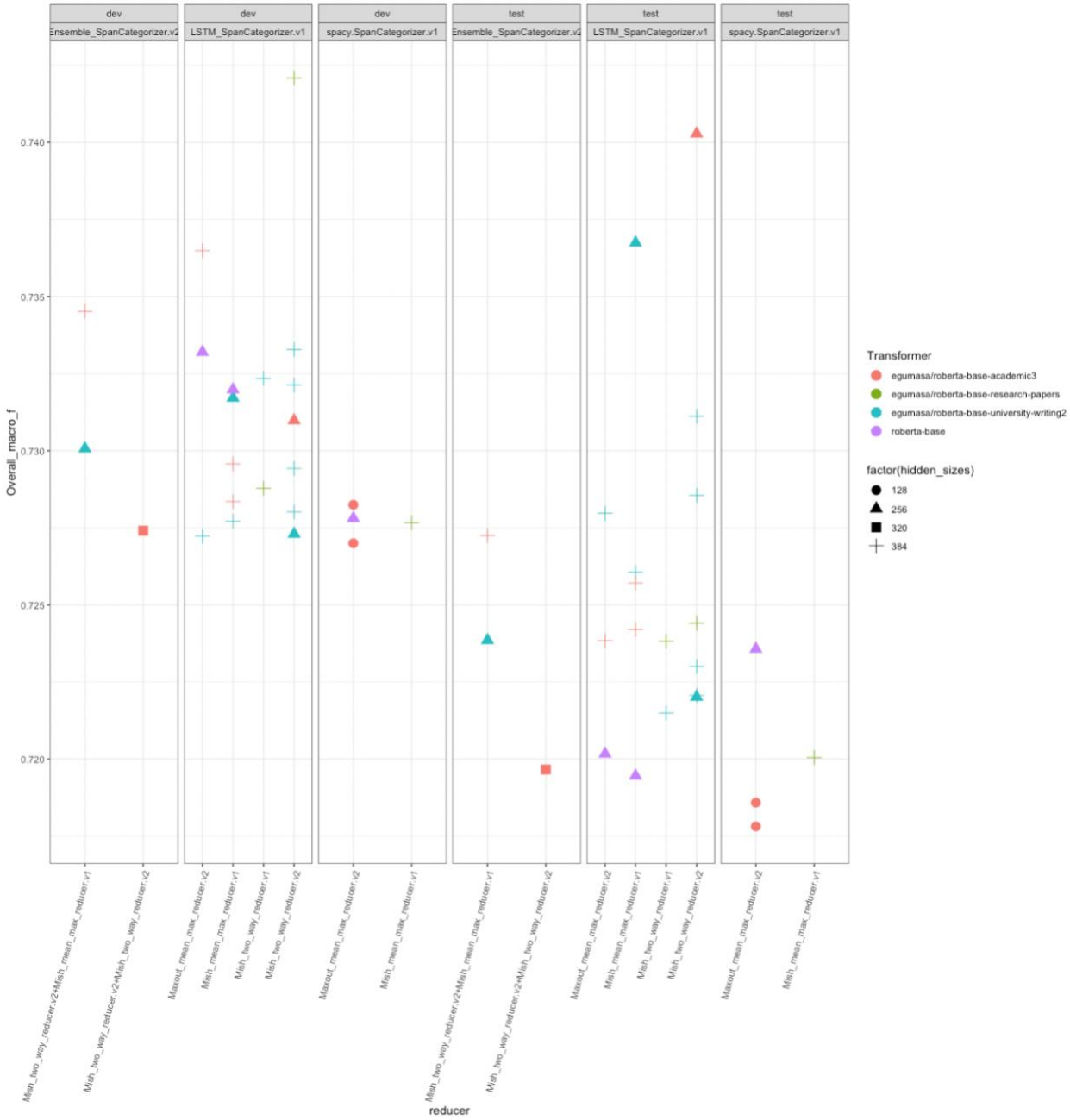
3.4.3 RQ3: What are the precision, recall, and F1 scores of the best-performing model?

The third goal of the current study was to determine the optimal combinations of architecture and hyperparameter settings. To this end, I retained models which performed in the top 25% on both the development and test sets and examined their hyperparameter combinations (Fig. 3.20). This resulted in 24 models in total. First, it appeared that Transformer+LSTM tended to achieve higher F1 scores overall, representing the majority in the retained model (17 of these, while there were 4 single-transformer and 3 dual-transformers models). A pairwise Fisher's exact test indicated that there was a statistically significant difference in the chance of the Transformer+LSTM model being included in the top 25 percentiles rather than the dual-Transformer models (adjusted p-value = .0207). However, this was not significant in a comparison of the Transformer+LSTM and single-Transformer models (adjusted p-value = .104). This provides additional support for the findings of RQ2, where the Transformer+LSTM model on average, tended to outperform the dual-Transformer model.

A follow-up qualitative analysis indicated that several Transformer+LSTM models appeared stable across the dataset tested, while there were general 0.5-to-1-point decreases in performances from the development to test sets in the other architectures. It is noteworthy that some models performed better on the test set, likely resulting from better representational learning due to the architecture. Further, the Transformer+LSTM architecture appeared to be robust against the selection of pre-trained Transformer models and non-linear activation functions. These results show the potential advantages of placing an additional LSTM layer after transformer embedding for the Engagement Annotation pipeline. This point is addressed further in the discussion below.

Figure 3.20

Top 25% scoring models for F1 scores and their hyperparameter settings.



Next, I conducted 5-fold CV on those models that showed high performance on both the development and test sets to investigate whether their architecture and hyperparameter settings were robust across the dataset. Table 3.12 shows the mean and min for each of their precision,

recall, and F1 scores, as well as kappa coefficients. The best-performing model was a variant of the Transformer+LSTM model, where the pre-trained model used the vanilla RoBERTa-base, scoring highest on Mean Precision (.754), F1 (.728), and Cohen's kappa (.689, which is considered moderate-to-substantial in the respective benchmarks; Landis & Koch, 1977; McHugh, 2012). The second-best model (according to mean F1 score) was a form of the dual-transformer model, which used a fine-tuned version of RoBERTa (averaged macro F1 = .721); however, this model did not score high on kappa coefficient (.658, second to last among the eight models). Based on the kappa coefficient, which also takes chance level agreement into account, the Transformer+LSTM models were the 1st to 4th best-performing systems.

The results for 5-fold CV further demonstrate the relative benefits of high-achieving models, which generally corroborates the findings for RQ2. First, the Transformer+LSTM architecture tended to score highest among the three architectures, particularly on Precision, F1, and Kappa. Second, the dual-Transformer scored highest on Recall, confirming the result for RQ2. An additional benefit of the dual-Transformer is that the ranges of F1 scores across 5-fold CV tended to be narrower, particularly in terms of their Minimum. This suggests that the model architecture is robust in that it produces stable results across five different held-out datasets. Finally, although the hypothesis testing for RQ2 showed that there were no significant differences between the single-Transformer and Transformer+LSTM architectures, the current result shows that the latter architecture may produce higher achievers.

Table 3.12
Results of 5-fold Cross-validation.

	5-fold CV (on held-out Test sets)											
	Macro P			Macro R			Macro F1			Kappa		
Architecture	*M*	*Min*	*Max*	*M*	*Min*	*Max*	*M*	*Min*	*Max*	*M*	*Min*	*Max*
Transformer												
(a) RoBERTa-base + Maxout (spaCy's default)	0.745	0.728	0.764	0.695	0.672	0.716	0.715	0.695	0.732	0.661	0.651	0.697
(b) RoBERTa-Academic + Mish	0.729	0.715	0.738	0.675	0.632	0.717	0.695	0.665	0.719	0.647	0.602	0.674
Transformer + LSTM												
(c) RoBERTa-base + LSTM + Mish two-way	**0.754**	0.734	**0.772**	0.710	0.670	0.734	**0.728**	0.696	**0.750**	**0.689**	0.634	**0.712**
(d) RoBERTa-Academic + LSTM + Mish two-way	0.752	**0.741**	0.763	0.691	0.667	0.715	0.715	0.696	0.726	0.678	0.648	0.682
(e) RoBERTa-Academic + LSTM + Mish two-way	0.743	0.710	0.763	0.704	0.666	0.726	0.719	0.696	0.736	0.674	0.643	0.697
(f) RoBERTa-Academic + LSTM + Mish two-way	0.747	0.735	0.770	0.695	0.679	0.719	0.716	**0.706**	0.725	0.665	**0.658**	0.686
Dual Transformer + LSTM												
(g) RoBERTa-Academic + Mish-two-way * 2	0.741	0.723	0.756	0.708	0.681	**0.737**	0.721	0.705	0.745	0.658	0.654	0.708
(h) RoBERTa-Academic + Mish-two-way * 2	0.718	0.691	0.744	**0.724**	**0.702**	0.733	0.718	**0.706**	0.732	0.664	0.635	0.692

Note. Columns which show the means of each metric are highlighted. Bold face indicates the highest score for each column.

Lastly, I examine the by-tag accuracy of high achievers because the end goal is to have a model that performs well across categories (for subsequent analysis in Studies 2 and 3). Table 3.13 presents by-category means and SDs based on the 5-fold CV results. Overall, I selected six high-performing models from Table 3.12 due to space limitations. This analysis indicated that the best-performing model (i.e., RoBERTa-base LSTM; [c]) outperformed the other models in six out of 10 categories. These categories were: ATTRIBUTION, CITATION, ENTERTAIN, MONOGLOSS, PROCLAIM, and SOURCES. It is noteworthy that the model performed with an F1 of over 70% on PROCLAIM and SOURCES, with which other models struggled. One weakness of the model (c) has to do with the ENDOPHORIC tag. On this tag, the model underperformed with an F1 of .7, and was placed second to last among the six models. Thus, there seemed to be a trade-off between categories.

Table 3.13

By-category average F1-scores via 5-fold Cross-validation.

Averages of 5-fold CV		Single-transformer		Transformer+LSTM						Dual-Transformer			
		(a) RoBERTa-base		(c) RoBERTa-base + LSTM		(d) RoBERTa-A + LSTM		(e) RoBERTa-A + LSTM		(g)		(h)	
	Human baseline	*M*	*Min*	*M*	*Min*	*M*	*Min*	*M*	*Min*	*M*	*Min*	*M*	*Min*
ATTRIBUTION	0.6	0.703	0.666	**0.756**	0.724	0.733	0.688	0.708	0.665	0.694	0.654	0.719	0.615
CITATION	0.95	0.905	0.887	**0.944**	0.895	0.928	0.893	0.906	0.890	0.919	0.879	0.911	0.901
COUNTER	0.85	0.852	0.739	0.857	0.820	0.867	0.806	0.871	0.825	0.877	0.857	**0.879**	0.857
DENY	0.87	0.856	0.822	0.855	0.785	0.867	0.822	0.877	0.797	**0.882**	0.819	0.842	0.852
ENDOPHORIC	0.62	0.724	0.600	0.660	0.545	0.637	0.530	0.701	0.678	0.731	0.605	**0.742**	0.685
ENTERTAIN	0.83	0.840	0.792	**0.845**	0.821	0.844	0.799	0.842	0.772	0.834	0.766	0.820	0.790
JUSTIFYING	0.82	**0.813**	0.777	0.784	0.648	0.808	0.781	0.802	0.752	0.808	0.748	0.796	0.761
MONOGLOSS	0.81	0.801	0.741	**0.821**	0.779	0.791	0.781	0.806	0.756	0.789	0.773	0.796	0.731
PROCLAIM	0.4	0.669	0.613	**0.741**	0.703	0.679	0.625	0.696	0.567	0.703	0.683	0.693	0.620
SOURCES	0.57	0.705	0.636	**0.751**	0.663	0.712	0.649	0.699	0.633	0.689	0.599	0.695	0.583
accuracy	0.72	0.703	0.689	0.723	0.673	0.709	0.686	0.711	0.696	0.706	0.677	0.701	0.692
macro avg	0.67	0.715	0.695	0.728	0.696	0.715	0.696	0.719	0.706	0.721	0.706	0.718	0.705
weighted avg	0.72	0.719	0.706	0.740	0.697	0.727	0.714	0.725	0.707	0.720	0.661	0.695	0.690

Note. An "O" label indicates empty tokens, and it is explicitly included in the table here because accuracy, macro average and weighted average, and kappa coefficients, included O tokens misclassified as labels; F1 scores under .7 are highlighted in red; F1 scores over .8 are highlighted in green.

3.5 Discussion

The aim of the current study was twofold—construction of the Engagement Discourse Treebank (EDT) and empirical evaluation of machine learning models that conduct automated Engagement resource analyses. To accomplish the first goal, a total of 126,000 words were manually annotated based on four interrelated layers—(a) clausal type annotation, (b) Engagement span and category, (c) Engagement hierarchy, and (d) supplementary rhetorical moves. Using annotations from (b), the Engagement layer, and four categories from (d), supplementary rhetorical moves, the second phase of the study concerned the training and evaluation of a series of machine learning models that identify spans of lexico-grammatical expressions and their categories. In what follows, I first provide a summary of the findings in answer to the three research questions presented in the chapter's introduction (see section 3.2.1). This summary of findings is then expanded in the remaining part for a discussion of implications and recommendations for further research. The demo version of the application can be freely accessed through HuggingFace Spaces at https://huggingface.co/spaces/egumasa/engagement-analyzer-demo.

3.5.1. Summary of findings

3.5.1.1 RQ1: What are the levels of intercoder agreement after annotators with linguistics backgrounds are trained on adapted schemes of the Engagement system?

The first research question concerned intercoder reliability estimates based on two sets of a human-annotated corpus. This agreement between human coders was also intended to set a benchmark against which the machine learning model's performance was evaluated. To derive intercoder reliability estimates, I used a subset of annotated data (a total of 35,601 tokens), which

were tagged by two coders completely blindly. As shown in Table 3.11, the Cohen's kappa coefficient between the two coders was .67; the macro F1 score (treating Annotator B as a "gold" tag) was also .67. F1 scores ranged from .40 to .95, indicating considerable degrees of difficulty in discerning rhetorical choices.

3.5.1.2 RQ2: What are the impacts of machine learning architecture selection and associated hyperparameters on precision, recall, and F1 scores?

The second research question concerned the effects of machine learning architecture and their associated hyperparameter choices on the accuracy of the Engagement Analyzer pipeline. To this end, I sketched out three different model architectures—(a) single-Transformer, (b) Transformer+LSTM, and (c) dual-Transformer+LSTM. I compared the relative performances of these three architectures and associated hyperparameter settings by conducting a random hyperparameter search (with random seed restarts). The resulting 205 models, based on a random hyperparameter search, were then submitted to a series of statistical analyses to see whether and how hyperparameter selections influenced the performance of the model. The results indicated that, overall, the Dual-Transformer+LSTM model underperformed the other two by a very small margin (Kappa = .69 for Dual-Transformer+LSTM as opposed to Kappa = .70 for the other two architectures). Additionally, the dual-transformer model tended to be Recall-focused at the expense of Precision, while the other two architectures scored slightly higher on Precision. In the F1 scores, there were statistically significant differences between the dual-transformer model and the other two models (F1 = .71 for Dual-Transformer+LSTM; F1 = .73 for Transformer+LSTM; F1 = .72 for single-Transformer). The dual-transformer models also tended to produce low-scoring outliers, suggesting that the architecture may be sensitive to other hyperparameter

settings. On average, the selection of a pre-trained model did not significantly influence the performance of the models when other things were equal.

3.5.1.3 RQ3: What are the precision, recall, and F1 scores of the best-performing pipeline of the Engagement Analyzer?

Given that the overarching goal of this study is to train a machine learning model to perform automatic Engagement analysis, RQ3 was interested in the best combinations of architectures and hyperparameters instead of average performance as examined in RQ2. In RQ3, I attempted to answer this question by focusing on the common hyperparameter settings of the top 25% of models on both the development and test sets. To this end, I conducted a series of 5-fold Cross-Validations with high-performing hyperparameter settings (Table 3.13). The results of this analysis indicated that Transformer+LSTM architectures tended to produce more high achievers (among the top 25% models, 70% for the Transformer+LSTM model). The best-performing NLP pipeline scored a macro F1 of .728 and Cohen's kappa of .689 (averaged over 5-fold CV data).

3.5.1.4 Organization of the remaining sections

The remainder of the chapter discusses several emerging themes from the current study, which are organized as follows. First, I will pick up the discussion regarding the challenges encountered during the corpus annotation project, particularly in relation to the distinctions between MONOGLOSSS and ATTRIBUTE, and MONOGLOSS and PRONOUNCE. I do this by drawing on notions of averral and attribution (e.g., Charles, 2006; Hunston, 2000), which are closely related to evaluative language but concepts distinct from the engagement system (see Hunston & Thompson, 2000). In the sections that follow, I will then focus on the findings from

the machine learning experiments. In particular, I will focus on the applicability of the current end-to-end approach to other tasks that might attract applied linguists. Finally, I also discuss two possible reasons for the performance boost by adding a Bi-LSTM layer on top of RoBERTa embedding. The chapter then concludes with the limitations of the current study and future research agendas for applied linguistics and NLP interfaces.

3.5.2. Challenges in annotating rhetorical moves

In this study, the annotation team faced multiple challenges in interpreting the Engagement framework and putting them into coherent annotation guidelines. Cohen's kappa coefficient of .67 (substantial; Landis & Koch, 1977; moderate; McHugh, 2012) between two trained coders implies multiple challenges in the identification of rhetorical moves and their lexico-grammatical realizations. This finding is reminiscent of Fuoli's (2018) claim that the descriptions and definitions in the extant literature on Engagement may be insufficient to categorize specific instances in discourse. To counteract these foreseeable problems, the current annotation project followed the stepwise annotation procedure proposed by Fuoli (2018), which puts emphasis on the replicability and transparency of the procedure. Specifically, the Method section of this study described the step-by-step procedure of the entire annotation project and its methodological decisions. The entire annotation manual is also made open source (https://egumasa.github.io/engagement-annotation-project/). Progressively tracking intercoder reliability four times during the entire annotation project would also contribute to maximizing intercoder agreement as well as transparency in the annotation process (Fuoli, 2018).

Despite these efforts, the study indeed re-discovered challenges in distinguishing some categories of Engagement—ENTERTAIN and PROCLAIM, MONOGLOSS and ENTERTAIN,

and MONOGLOSS and ATTRIBUTION. Although this study is not intended to refine the Engagement system, these challenges may still offer some insights for future research (cf. Creswell & Poth, 2016). The remainder of this section outlines two important decisions that the annotation team made in the Engagement guidelines (https://egumasa.github.io/engagement-annotation-project/) to enhance the intercoder reliability of the analysis as well as the credibility of their discourse interpretation.

One of the most problematic is the category of PRONOUNCE. Martin & White (2005) describe PRONOUNCE thus:

> The category of pronounce covers formulations which involve authorial emphases or explicit authorial interventions or interpolations. For example: *I contend*..., *The facts of the matter are* that..., *The truth of the matter is* that..., *We can only conclude* that..., *You must agree* that.., intensifiers with clausal scope such as *really*, *indeed*, etc. and, in speech, appropriately placed stress (e.g., *The level of tolerance IS the result of government intervention*). (Martin & White, 2005, p. 127)

What is significant from this description is that PRONOUNCE has to do with utterances involving "authorial emphases, interventions or interpolations" (p.127). A few other examples that Martin and White (2005) use to illustrate the idea of PRONOUNCE include:

- *It is absolutely clear to me* that…
- *We have to remember* that…
- What *really* differentiates cool from worm couples is…

One of the struggles the annotation team faced was the treatment of commonly occurring clausal expressions in academic discourse such as "*it is important*", "*it is essential*", or "*it is evident*" (called extrapolated that-clauses followed by an adjectival complement ; Biber et al., 2021). In the SFL tradition (e.g., Halliday & Matthiessen, 2014), these extrapolated that-clause constructions are often considered a means to express evaluation objectively (Halliday &

Matthiessen, 2014, p. 688; Martin & White, 2005) by hiding the agentive role of the writer in the immediate context (by way of an impersonal construction). This is also referred to as the interpersonal metaphor (Halliday & Matthiessen, 2014), where modality expressions are realized with incongruent grammatical structures (i.e., an It is X that-clause).

After discussing the issue with the annotation team and reviewing the literature on evaluative language, I decided that the categorization of extrapolated that-clauses would primarily be motivated by the lexical semantics (and by extension, functional potentials) of the adjectival complements which control the that-clauses. For example, when the controlling adjectives are "important" or "essential", there is good reason to treat them under MONOGLOSS. Based on the dialogic potential of the expressions "it is important" or "it is inevitable", these would be considered as undialogized utterances (ignoring the dialogic potential in the immediate discourse). This undialogized nature of utterances is supported by the fact that one can add ENTERTAIN resources to show recognition of this dialogic nature of discourse (e.g., It *seems* important; It *can* be essential). Another reason is that the extrapolated that-clauses in "it is important" and "it is essential" apparently lack the author's explicit commitment to the proposition compared to prototypical examples of PRONOUNCE, such as "I *contend*" or "I *conclude*". Thus, a seemingly un-dialogized, non-explicit comment adjective ("important", "critical", "necessary") would fit well with the prototype description of MONOGLOSS in Martin and White (2005). On the other hand, variants of extrapolated that-clause constructions can be categorized under PRONOUNCE when the adjectives reveal the writer's epistemic commitment to the proposition introduced in the that-clause (e.g., "*it is clear*" and "*it is evident*"). This decision would also be supported by the fact that their derived adverbial forms are introduced as prototypical PRONOUNCE-enacting resources ("clearly" and "evidently"; Martin & White,

2005). To the best of my knowledge, there is at least one previous research example in the Engagement literature that included "*it is evident*" as an example of PRONOUNCE (Chang & Schleppegrell, 2011), although I was not able to identify mentions of this particular construction in the original Martin and White (2005) monograph. Such extrapolation of the original Martin and White (2005) volume may invite a potential pushback from SFL analysts, but a comprehensive analysis of this construction is beyond the scope of the current study.

A second major theme that emerged from the annotation project was the distinctions among MONOGLOSS, PRONOUNCE, ATTRIBUTE, and ENDORSE (also identified by Fuoli, 2018). Several possible sources of disagreement were identified in the annotation project. The first of these, which was identified at an early stage of the project and thus motivated the addition of a supplementary tag, had to do with discourse moves that are typical of academic writing—namely, references to Tables, Figures, and other information that is derived from the writer's own research. The following two excerpts from the EDT illustrate this type of discourse move:

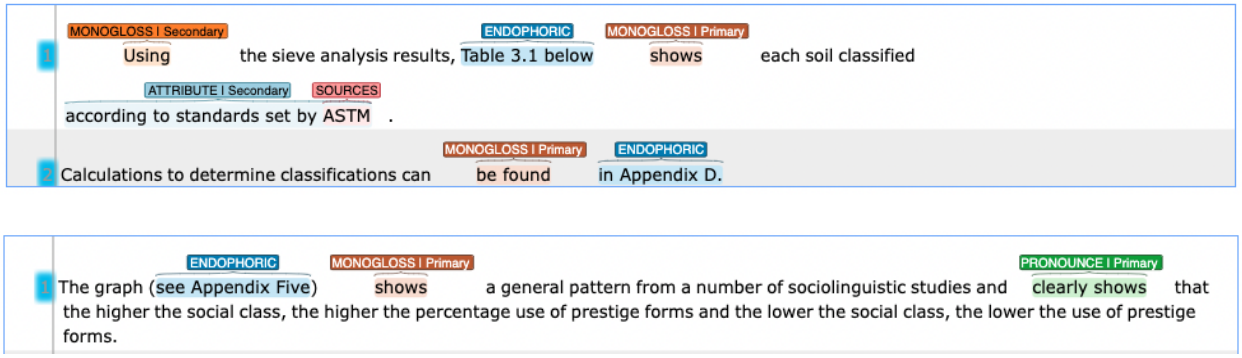
- *Using the sieve analysis results, Table 3.1 below shows each soil classified according to standards set by ASTM. Calculations to determine classifications can be found in Appendix D.* (EDT ID: CEE.G0.01.3_11_2)
- *The graph (see Appendix Five) shows a general pattern from a number of sociolinguistics studies and clearly shows that the higher the social class, the higher the percentage use of prestige forms and the lower the social class, the lower the use of prestige forms.* (EDT ID: 6126d_s4.8; p12.1)

In Martin and White (2005) and other research on Engagement, the communication verb "show" is often presented as a prototypical resource for ENDORSE. Martin and White (2005) describe ENDORSE as "formulations by which propositions sourced to external sources are construed by

the authorial voice as correct, valid, undeniable or otherwise maximally warrantable" (p. 126). They list several factive verbs (as described in Biber et al., 1999) such as *show*, *prove*, *demonstrate*, *find* and *point out* as prototypical ENDORSE resources (p.126). According to the definition of ENDORSE, however, these verbs do not always enact ENDORSE but would probably have more characteristics of MONOGLOSS or PRONOUNCE. Our new heuristics, explained below, resulted in the annotation of the communication verb *show* and its co-textual features as shown in Figure 3.21.

Figure 3.21

Illustrative examples of annotated data which include communication verb "show"



According to our new heuristics, communication verbs such as *show*, *indicate*, among others, should be classified by considering the exact communicative functions of these items in the immediate discourse. For instance, in the first example, we consider that *Table 3.1 below* is a text-internal mention, not a text-external mention, which would fall out of the definition of ATTRIBUTE or ENDORSE in the original Martin and White (2005) definition. In our reading, *show* in these contexts does not have an argumentative function but instead signals where specific information is located in the piece of writing. The latter reading then implies that

MONOGLOSS would be an appropriate category for this particular pattern. Likewise, the factive verb *found* (line 2; Figure 3.21) is considered MONOGLOSS not ENDORSE. Also, I argue that the modal auxiliary *can* (i.e., *Calculations to determine classifications* ***can*** *be found in Appendix D*. line2, Figure 3.21) does not have an ENTERTAIN value in the specific context because of its focus on the textual function of the discourse (or directive) and is more characteristic of dynamic modality instead of epistemic (Palmer, 2001). In the second excerpt, we identified that the two instances of *show* are being used for different functions. The first use is prototypical of the ENDOPHORIC + MONOGLOSS mentioned above. The second use is more prototypical of a PRONOUNCE reading on the ground that we can recognize the explicit writer's commitment to the proposition presented in the that-complement clause. We did not consider this case as ENDORSE because, in our new heuristics, the wording within the minimal context suggests that the source of information, the graph in Appendix Five, can be taken as the writer's own piece of work, not work by another writer.

One possible reason why we encountered an issue with ENDORSE, MONOGLOSS, and PRONOUNCE is likely because the original Martin and White (2005) framework for Appraisal analysis focused on the analysis of media discourse (e.g., newspaper articles, public speech). In that context, lexical items listed as typical ENDORSE items such as *show*, *indicate*, *demonstrate*, would be predominantly for ENDORSE purposes—i.e., to draw on a third person's argument to make their own case credible. The new heuristics, an adaptation of Martin and White's (2005) original definition of the category of rhetorical functions of academic discourse, will be able to capture more instances of both interpersonal and textual meaning-making processes. Although proposing a new discourse framework is beyond the current study, the insights gained from the current annotation project may inform future refinement of the discourse analytic framework to

analyze Engagement in academic discourse and evaluative language by extension (see Fuoli, 2018 for a role of annotation as theory building process). Also, making the annotated corpus available to the public will make it possible for researchers to closely examine the validity of previous annotations, thus stimulating collaboration among researchers. Such cross-disciplinary collaborations will make EDT a more rigorous manually annotated dataset in the future.

3.5.3 Leveraging state-of-the-art NLP models for custom linguistic analyses in applied linguistics

The current study has demonstrated that a probabilistic approach to identifying rhetorical strategies of Academic English writing provides a moderate degree of accuracy (Cohen's kappa = .689; macro F1 = .728; micro F1 = .74) when trained on the adjudicated human annotation. Comparing the best-performing model to the intercoder agreement reported in RQ1 (kappa = .67; macro F1 = .67), the automated NLP pipeline is comparable to (or even outperforms) trained human coders. These findings have clear implications not only for the analysis of evaluative language (Hunston & Thompson, 2000; Hyland, 2005a; Martin & White, 2005) but also corpus-based register analysis (Biber, Conrad, Reppen, et al., 2004; e.g., Biber, 2006a) and, by extension, the assessment of written performance using corpus linguistics and/or natural language processing techniques (Attali, 2007; Bax et al., 2019; Yoon, 2017b).

The current probabilistic approach would be a viable (and probably better) alternative to the analysis of stance-taking features in corpus-based register analysis (Biber, 1988; Biber, Conrad, Reppen, et al., 2004). As argued previously, most of the currently available corpus linguistic tools analyze the features of stance (Bax et al., 2019; Biber, 2006a; Yoon, 2017b) through dictionary-based pattern matching and/or regular expression search. This approach often

follows a two-step procedure, where the researcher compiles a list of "functionally important" expressions through quantitative (plus qualitative) analyses of large-scale reference corpora, followed by pattern matching search on a given text using a compiled list. This approach, particularly the first step of this method, has been shown to be useful as it has revealed lexico-grammatical patterning, which tends to be associated with rhetorical functions in particular genres of spoken and written texts (lexical bundles; e.g., Biber, Conrad, & Cortes, 2004). The second step—the analysis of subsequent texts based on the identified lexical strings—does not consider the exact rhetorical functions for which an item is being used in a text. This is not entirely problematic when the item in question is predominantly used for a distinct rhetorical function (e.g., *in other words*). This mono-functional assumption is problematic, however, because, as many researchers note (Hunston & Thompson, 2000; Hyland, 2005a; Martin & White, 2005, 2005), a myriad of stance-taking expressions are interpreted with their co-textual information in mind (e.g., the communication verb *suggest* can be used either to make an attribution or present a hedged statement). For this reason, I propose that the present probabilistic machine learning approach can overcome this limitation of the existing corpus-based approaches, which tend to rely on rule-based pattern matching (e.g., Bax et al., 2019; Biber, 2006a; Yoon, 2017a).

A clear benefit of the current machine learning approach is that is allows the analysis of large amounts of texts at scale. The previous discourse analytic approach to stance (Hyland, 2005a; Martin & White, 2005) focused on a smaller number of texts to provide a detailed account of the ways in which rhetorical features are used. The current approach expands the scope of research by allowing the (semi-)automatic analysis of these features, which in turn allows large-scale automated discourse analysis. Given this extension, it should be possible to

confirm insights gained in previous qualitative research with large quantities of data to see if patterns are generalizable across specific groups of writers, disciplines, and genres of writing, to name but a few relevant variables. This is the scope of Study 2 of the current dissertation (see Chapter 4).

The automated analysis of fine-grained rhetorical discourse features introduced in this chapter also has implications for automated writing assessment and second language writing research. As an increasing number of language assessment researchers incorporate insights gained from corpus-based register analysis (e.g., Kyle et al., 2022; LaFlair & Staples, 2017; Staples et al., 2018), it would be beneficial to explore the potential use of the Engagement Analyzer to explain certain facets of second language writing performance. If rhetorical features measured through the Engagement Analyzer are of any use in explaining additional variance in the assessed performances of second language writing, it will indicate that the Engagement Analyzer can contribute to the increasing construct coverage in writing skills assessment. This is the scope of Study 3 of the current dissertation (see Chapter 5).

Beyond the immediate context of the study, the current study has showcased the potential use of a state-of-the-art machine learning approach to construct an end-to-end NLP pipeline to conduct custom linguistic analysis. Specifically, I used a variant of a pre-trained language model (i.e., RoBERTa; Liu et al., 2019) to analyze theoretically important constructs in research on English for Academic Purposes (EAP). The overall steps of the research can be summarized as follows:

Step 1: Defining linguistic or discourse constructs and in-domain text

Step 2: Annotating corpora sampled from a defined domain

Step 3: Training ML model(s)

Step 4: Evaluating of accuracy of ML models against (human) benchmark

(Step 5): Developing application, including user-interface and demo versions

In this process, it is the job of applied linguists to define the linguistic or discourse constructs that are necessary to advance the field (see Lu, 2021). For example, in this study, I chose to operationalize the Engagement framework (Martin & White, 2005) as a theory to explain interpersonal evaluative language; however, it is possible to draw on other frameworks to annotate data and train an automated NLP pipeline, such as the meta-discourse framework from Hyland (2005).

More research is needed to make this approach more accessible to applied linguists. For instance, it is still unclear how much data is enough to achieve acceptable accuracy in end-to-end NLP models. Although this depends on the complexity of the task and available resources (how large a network one can train), this information would be informative to applied linguists who would like to automate the analysis of a given linguistic feature for their respective research domain. More collaboration between applied linguists and computational linguists will raise more questions to be answered to address specific analytic questions, given the epistemologies and methodologies of respective research domains.

3.5.4 Potential benefits of extra contextual information?—RoBERTa+Bi-LSTM

The finding of the current study also has implications to computational linguistics/natural language processing research. The results of this study suggest that adding a one-layer Bidirectional LSTM on top of RoBERTa embeddings may enhance model performance. Although this kind of RoBERTa+Bi-LSTM architecture is not uncommon in span detection tasks, studies have reported mixed findings regarding its benefits (Gu et al., 2022; Papay et al.,

2020; Zhu et al., 2021). For this reason, the potential mechanisms which led to slightly better performance in the RoBERTa+Bi-LSTM architecture merit the discussion conducted below. These potential gains may have stemmed from two characteristics, which may exert effects in combination.

The first possible reason for the gains from an additional Bi-LSTM layer is that they may provide task-specific sequential information for span layers. This seems to be a consensual intuition regarding the benefits of an additional Bi-LSTM layer in a span detection/ categorization task (Gu et al., 2022; Zhu et al., 2021). In this thinking, contextually aware token embeddings by the Transformer model can be further refined by the gating mechanisms in Bi-LSTM (Hochreiter & Schmidhuber, 1997). Although the LSTM architecture is much simpler (and importantly without the multi-headed self-attention mechanisms like in the Transformer architecture), such sequential information may be particularly helpful for edge detection, which is critical in the current span-detection architecture. That is to say, given that the current span suggester was greedy in that it recommends all possible dependency subtrees and n-grams, additional information regarding surrounding tokens may have been immediately of use (e.g., controlling the amount of information from the surrounding context to be considered). Although I allowed the RoBERTa embeddings to be fine-tuned during training, an additional layer of LSTM may have helped to determine such task-specific weights in addition to the contextual embedding of running tokens alone. Indeed, this finding may partly explain the findings of Gu et al. (2022), who found mixed benefits from adding Bi-LSTM. In their implementation, Bi-LSTM added on top of a RoBERTa embedding boosted the performance on some datasets when their architecture was similar to the current span suggester approach (i.e., Span Emulation, Gu et al., 2022; see also K. Lee et al., 2017), although there was not much benefit when the task was

formulated with regard to token-by-token sequence labels (in an I, O, B format). Thus, while Gu et al. (2022) concluded that the benefits of this approach were limited (corroborating the limited gains found in Zhu et al., 2021), the results of the current study suggest that the additional LSTM layer may well stabilize predictions depending on the specific set-up of the prediction task (edge prediction and/or span enumeration as opposed to token-by-token tagging; Gu et al., 2022).[3]

A second possible reason for the current gain may have to do with the fact that the hidden dimensions of the Bi-LSTM are set at 200. That is, the gains may not be solely due to richer contextual information, but also because the current Bi-LSTM layer acted as a dimensional reduction of the RoBERTa embeddings, allowing for a subsequent span candidate pooling layer to further condense the information. If this is the explanation, the dimensional reduction explanation may be ruled out by using a Bi-LSTM with the same hidden dimensions with a RoBERTa embedding; however, it appears that the previous work does implement smaller hidden sizes for the LSTM layer (200, Gu et al., 2022; 256, Z. Jiang et al., 2020). This may merit further study.

3.5.6 Limitations and directions for further research

Before concluding the study, a few important limitations should be noted for future studies. The first limitation concerns the fact that the Engagement Discourse Treebank (EDT) took a minimal context approach in sampling, rather than considering a whole document as a unit for sampling. The practical considerations of copyright issues of original documents as well as

[3] One possible advantage of the span enumeration (the architecture used in this study), compared to token-by-token sequence labels as in Named Entity Recognition task, include that it allows the tagging of nested, overlapping spans of multiple categories, where the sequence labeling task (i.e., formulated as tagging) may not be readily extendable to multiple overlapping span labeling problem (see Gu et al., 2022).

the necessity to represent as many writing styles as possible resulted in this necessary compromise in the amount of contextual information available in both the annotation and training of the machine learning component. Thus, the current version of the Treebank should be regarded as an initial attempt to expand the scope of register-based corpus linguistic tools, which predominantly take rule-based approaches (i.e., list-matching, regular expressions). The results of this study show that an end-to-end approach could be a promising direction to expand corpus studies which attempt to help the annotation of a given language sample for any theoretically interesting constructs. Therefore, the next generation of the EDT should consider a whole document as a unit of analysis so as to be able to take full advantage of available contextual information in the manual annotation process.

Relatedly, in this study, I ruled out the possibility of using Transformer models for longer sequences such as Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), and Transformer-XL (Dai et al., 2019). However, given the nature of the task, future iterations of this project should consider these pre-trained models with an expanded document length. Another future direction related to the machine learning architecture is to develop a custom span suggester that is less greedy. As one may be able to define the span in terms of clear grammatical boundaries, a rule-based approach with dependency representation may reduce the false positives significantly. As mentioned already, a neural approach to predict the span was also tested during the early experiment (similar to span prediction in Gu et al., 2022), but this approach limited the recall at the span suggestion layer (Recall = .7), hurting the overall accuracy of the end-to-end results. With updates in the neural architectures, the model-based prediction of span suggester may be optimized for recall, which then may increase the accuracy of the overall pipeline. Finally, the three networks tested in this study were relatively complex, and one may wonder if a

simpler architecture may work equally as the best-performing model. Through the current experiment, however, the best-performing model required additional LSTM layer on top of the spaCy baseline span categorizer. Thus, based on the current experiment, a similar performance to the best-performing model with a simpler architecture might be difficult to obtain. These topics merit further research.

Next, although the current version of EDT amounts to 126,411 tokens with 4,688 sentences in total, this size is not sufficient to cover the ranges of possible linguistic realizations of some minority categories. For example, in the entire annotated corpus, among the least frequent engagement strategies were CONCUR (k =127) and ENDORSE (k =147). Given such small numbers of occurrences, it was necessary in the current study to merge the two categories with their closest relevant categories (i.e., CONCUR and PRONOUNCE, and ATTRIBUTE and ENDORSE). While merging adjacent categories have been done in previous research on Engagement to enhance intercoder reliability (Fuoli, 2018; Fuoli & Hommerberg, 2015), it inevitably leads to a loss of granularity in the analysis. For this reason, it would be beneficial to increase the size of manually annotated corpora and update ML models. In so doing, it may be beneficial to take a weak supervision approach, where the current best-performing models are used to provide baseline annotation.

Finally, although the current study improved the levels of intercoder agreement in identifying and categorizing Engagement strategies compared to preexisting benchmark by Read and Carroll (2012), who annotated the entire Appraisal network, the absolute levels of agreement (F1 = .67 between the two coders) was far from satisfactory. One possibility, which echoes Fuoli's (2018) argument, is that the category definitions and descriptions given by Martin & White (2005) are insufficient to rule out subjective interpretations of the coders as specific

realizations may well be context-dependent affected by the consideration of genres and disciplinary conventions, among others. That said, although the current annotation team (lead by a Ph.D. candidate in linguistics and two undergraduate students in linguistics) may have had a reasonable conceptual grasp of prototypical realization of engagement strategies based on excerpts and qualitative analysis presented in published materials (Chang & Schleppegrell, 2011; Lancaster, 2014; Martin & White, 2005; Ryshina-Pankova, 2014; White, 2003; Wu, 2007; Xu & Nesi, 2019), it is unlikely that they have seen the every possible way in which engagement meaning can surface in different genres and disciplines included in the annotation data randomly sampled from various disciplinary areas within the academic domains. Therefore, the task of annotating engagement strategies may require helps of more experienced discourse analysts and/or domain experts who knows the disciplinary conventions. Based on Fuoli's (2018) principal recommendations, more research is needed to come up with strategies to improve the transparency and accountability of linguistic annotation for discourse-oriented applied NLP research.

3.6 Chapter Conclusion

This chapter has described the design features of the Engagement Discourse Treebank (EDT) and Engagement Analyzer. It has also reported annotation projects and a series of machine learning experiments to train an end-to-end NLP pipeline to conduct the automatic analysis of stance-taking features of academic English writing, which draws on the framework of Engagement (Martin & White, 2005) and additional supplementary discourse features (Hyland, 2005a). The Engagement Analyzer is intended to provide a means to overcome the dilemma faced by two lines of research on stance in academic written English. The ability of the Engagement Analyzer

to reach human-level agreement on the identification of stance-taking features will enhance the ability to conduct large-scale studies on register variations across disciplines and genres, as well as the application of a discourse analytic framework to second language assessment. Each of these topics form the scope of Studies 2 and 3. The demo version of the application can be accessed through HuggingFace Space at https://huggingface.co/spaces/egumasa/engagement-analyzer-demo.

CHAPTER 4

STUDY 2: Register Variations of Engagement Strategies Across University Written Assignments

4.1 Chapter Overview

In Chapter 3, I reported the development and evaluation of the Engagement Analyzer, an end-to-end machine learning system that undertakes automated engagement resource analysis (Martin & White, 2005) of academic written English. The findings suggest that the accuracy of the system (evaluated through Precision, Recall, F1 scores, and Cohen's Kappa) is comparable to trained human coders. Given this finding, Study 2 aims to apply a developed system to describe register variations of university written assignments using a large-scale corpus of this domain—the British Academic Written English corpus (BAWE corpus; Alsop & Nesi, 2009).

4.2 Research Questions

RQ How does the use of engagement strategies in written university assignments vary across:

a. disciplines,
b. genre families,
c. course levels,
d. assignment grades, and
e. the first language of the writer?

4.3 Method

4.3.1 British Academic Written English (BAWE) corpus

The data for this study came from the BAWE corpus (Alsop & Nesi, 2009), which contains 2,761 authentic university assignments submitted to four universities in the UK from 2005 to 2007 (see Table 4.1). The BAWE corpus is one of the most comprehensive corpora of university writing, distributed across 30 disciplines, 13 genre families, and four course levels (from freshman to master's level post-graduate courses). Most of the submitted assignments were considered "successful" as they were above Distinction ("D") and Merit ("M"), roughly corresponding to letter grades A and B respectively (Nesi & Gardner, 2012, p. 9). Below, I will outline relevant metadata for this study—disciplines, genre families, and writer-related information in the corpus. In the current study, I used a subset of 2,685 assignments in the analysis for the reasons mentioned below (see section 4.3.1.3).

4.3.1.1 Author-related information—L1, educational background, gender, and course

The writers' self-reported demographic information is available in the BAWE corpus (Alsop & Nesi, 2009). This includes their L1 (e.g., Arabic, Chinese Cantonese, English, Spanish), whether they graduated from secondary school in the UK or overseas, their gender, and their year of birth. Author-related metadata also include courses (roughly equivalent to a program or degree in the US system, e.g., BA in Archeology, MSc in Chemistry).

4.3.1.2 Modules, Disciplines, and Disciplinary Groups

Each assignment in the BAWE corpus can be identified in terms of the module title it was submitted for (roughly equivalent to a course in the US university system), discipline, and

disciplinary group. Each module is classified into 30 disciplines, which in turn belong to four disciplinary groups—Arts and Humanities (AH), Life Sciences (LS), Physical Sciences (PS), or Social Sciences (SS). In other words, these three variables create a nesting structure in the dataset (i.e., disciplinary groups > disciplines > modules).

Table 4.1
Composition of the BAWE corpus.

Disciplinary group	Disciplines	Level 1	Level 2	Level 3	Level 4	Sum
Arts and Humanities (AH)	Linguistics, English, Philosophy, History, Classics, Archaeology, Comparative American Studies, Other	231	225	160	77	693
Life Sciences (LS)	Biology, Agriculture, Food Sciences, Psychology, Health, Medicine	172	183	106	193	654
Physical Sciences (PS)	Engineering, Chemistry, Computer Science, Physics, Mathematics, Meteorology, Cybernetics & Electronics, Planning, Architecture	181	146	156	95	578
Social Sciences (SS)	Business, Law, Sociology, Politics, Economics, Hospitality Leisure & Tourism Management, Anthropology, Publishing	203	196	159	202	760
Sum		787	750	581	567	2,685

4.3.1.3 Genre family

Each assignment in the BAWE corpus is manually classified into one or more genre families (Table 4.2). Rather than adopting the contributors' labels, all the texts in the BAWE corpus were closely examined by the corpus-building team through genre analysis (Alsop & Nesi, 2009; Nesi & Gardner, 2018). This bottom-up classification of the communicative, situational features of the assignments allowed the research team to propose a total of 13 genre families, which in turn belong to five broad categories of macro-social and educational purposes (Nesi & Gardner, 2012, 2018). Table 4.2 (adapted from Nesi & Gardner, 2018) shows the five broad social purposes of written assignments, 13 genre families, and example genres. In the BAWE corpus, a small number of assignments (66 out of 2,761) were classified into multiple genre families because they consisted of multiple parts. For example, an Engineering business management course assignment can include a case study discussing a company's performance or future direction, followed by a narrative recount of the case study process (e.g., BAWE document ID: 0090b, 0340b). For simplicity of analysis, I focus on single-genre assignments in this study.

4.3.1.4 Levels and Grades

Other relevant metadata concern the level of the module and assignment grades. The level is a four-level ordered category, where a value from 1 to 3 indicates the course level in undergraduate programs, and 4 represents a master-level post-graduate module. The grade is a three-level factor comprising Merit (60-69%), Distinction (70-100%), or Unknown.

Table 4.2
Summary of genre family classification and example genres in the BAWE corpus.

Social purpose	Genre family	Level 1	Level 2	Level 3	Level 4	Sum
Demonstrating knowledge and understanding	exercise	28	20	27	27	102
	explanation	72	54	33	26	185
Developing powers of independent reasoning	critique	75	78	67	87	307
	essay	398	357	263	184	1,202
Building research skills	literature survey	10	6	7	9	32
	methodology recount	106	114	43	60	323
	research report	7	16	22	16	61
Preparing for professional practice	case study	26	30	34	98	188
	design specification	24	19	35	11	89
	problem question	12	18	5	2	37
	proposal	10	18	8	34	70
Writing for oneself and others	empathy writing	4	2	17	5	28
	narrative recount	15	18	20	8	61
	Sum	787	750	581	567	2,685

4.3.2 Engagement Analysis

A total of 2,685 single-genre assignments from the BAWE corpus, which excludes 66 multi-genre assignments from the entire corpus, were analyzed in this study. The identification of engagement strategies was conducted automatically using the Engagement Analyzer (Chapter 3).

As discussed in Chapter 3, the Engagement Analyzer is an end-to-end machine learning system trained on a human-annotated corpus, the Engagement Discourse Treebank (EDT; see Chapter 3). The accuracy of machine learning systems tended to outperform the benchmark of two human annotators, who majored in Linguistics and were trained over ten weeks (over 50 hours of working time). However, it is important to acknowledge that the accuracy of a single model was still moderate-to-substantial (Cohen's Kappa = .72). For this reason, the current study employed four separately trained models from Chapter 3 to gain from the relative strengths of each of these models. This process is analogous to hiring four different human coders for data analysis. Table 4.3 lists the performances of the four models on the test sets for each one.

As can be seen from Table 4.3, three LSTM-based models (models [1]-[3]) and one Dual-Transformer model (model [4]) were considered. This choice was motivated by the relative strengths of these models and the distribution of the held-out datasets they were trained on. Three versions of the best-performing LSTM model, with the same architecture and hyperparameter settings but trained on three versions of held-out datasets, were included. Their by-tag performances were excellent, except for the ENDOPHORIC tag by model (2). To mitigate this weakness, I included one version of the Dual-Transformer model, which performed exceptionally well on the ENDOPHORIC tag. It is important to stress that although the overall accuracy figures are similar, each model has relative strengths. It is hoped that this demonstrates their strengths and the fact that they were trained and tested on different held-out datasets boosts the reliability of the overall analysis. The variabilities in the analyses by the four different models were statistically controlled for through mixed-effect (or multilevel) modeling (see section 4.3.4, Statistical analysis).

Table 4.3

Summary of models' performances used to identify engagement strategies.

	LSTM model trained on Fold1			LSTM model trained on Fold3			LSTM model trained on Fold5			Dual Transformer model trained on Fold4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ATTRIBUTION	0.762	0.689	0.724	0.841	0.745	**0.791**	0.813	0.696	0.750	0.790	0.791	*0.790*
CITATION	0.965	0.976	**0.970**	0.912	0.932	0.922	0.948	0.964	*0.956*	0.904	0.899	0.901
COUNTER	0.836	0.963	**0.895**	0.903	0.836	0.868	0.836	0.888	0.861	0.921	0.854	*0.886*
DENY	0.950	0.884	**0.916**	0.830	0.835	0.833	0.863	0.838	0.850	0.854	0.886	*0.870*
ENDOPHORIC	0.703	0.723	0.713	0.528	0.563	0.545	0.900	0.619	0.734	0.858	0.837	**0.848**
ENTERTAIN	0.890	0.779	0.831	0.869	0.868	**0.869**	0.865	0.873	**0.869**	0.832	0.869	*0.850*
JUSTIFYING	0.862	0.821	*0.841*	0.957	0.758	**0.846**	0.819	0.790	0.804	0.785	0.825	0.805
MONOGLOSS	0.822	0.845	0.833	0.867	0.830	**0.848**	0.856	0.818	*0.837*	0.871	0.735	0.797
PROCLAIM	0.873	0.658	*0.750*	0.837	0.765	**0.799**	0.818	0.667	0.735	0.730	0.652	0.689
SOURCES	0.830	0.736	**0.780**	0.744	0.800	*0.771*	0.741	0.803	0.770	0.770	0.755	0.763
_	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
accuracy			0.741			*0.743*			**0.742**			0.741
macro avg	0.772	0.734	**0.750**	0.753	0.721	0.736	0.769	0.723	0.742	0.756	0.737	*0.745*
weighted avg	0.786	0.741	*0.761*	0.791	0.743	**0.765**	0.775	0.742	0.756	0.761	0.741	0.749
kappa			0.710			**0.712**			*0.711*			0.708

Note. F1 columns are highlighted for readability; best F1 scores (by category) are bolded; second-best F1 scores (by category) are italicized; Scores under .7 are highlighted in red and scores above .8 are highlighted in green.

4.3.3 Engagement measures

In this study, I used an assignment (i.e., individual BAWE corpus file) as a unit of analysis. Before the analysis, all files were preprocessed to exclude the following sections based on XML tags in the corpus files: (a) references or bibliographies, (b) mathematical formulae not embedded in the prose text, (c) list-like elements, (d) figure and table captions, and (e) section headings. This preprocessing allowed the analysis to focus on the writer's main argument.

4.3.3.1 Frequency counts

To answer the research question, four separate datasets were created. Each time, the preprocessed corpus files were submitted to one version of the Engagement Analyzer, which returned item-level information, including lexico-grammatical items, their grammatical realization, and engagement category (one of 10 labels). Since I used an assignment (or corpus document) as the unit of analysis in this study, the number of Engagement strategies was tallied for each document in each of the four datasets. Next, the four datasets were concatenated for subsequent analysis. This resulted in four separate rows of occurrence counts for each assignment (from each model run), with ten columns representing each Engagement category and additional columns for assignment-level meta-data, such as student ID, discipline, genre family, length of the main text in words, etc.

Before the main analysis, the internal consistency of the assignment level frequency counts of engagement strategies across four versions of the Engagement Analyzer was examined. Since I was interested in assignment-level aggregated frequency counts for the primary analysis, a series of Spearman rank-order correlations and Mean Inter-item Correlations for each of the ten Engagement categories were computed (Table 4.4). This analysis revealed that the internal

consistencies of document-level frequency counts across models were strong (all pairwise correlations were above .871). This suggests that, when aggregated at the document level, the performance of the Engagement Analyzer may be more robust than might be deduced from item-level Kappa and F1 statistics.

Table 4.4
Reliability of document level frequency counts across four versions of the Engagement Analyzer.

Categories	Inter-item Correlation (rho)	Pairwise Spearman correlations					
		(1) - (2)	(1) - (3)	(1) - (4)	(2) - (3)	(2) - (4)	(3) - (4)
ATTRIBUTION	0.982	0.979	0.980	0.977	0.981	0.977	0.978
CITATION	0.957	0.980	0.968	0.957	0.984	0.973	0.971
COUNTER	0.991	0.990	0.991	0.991	0.991	0.992	0.992
DENY	0.981	0.984	0.983	0.977	0.983	0.976	0.979
ENDOPHORIC	0.917	*0.896*	0.900	*0.871*	*0.893*	*0.895*	*0.888*
ENTERTAIN	0.990	0.986	0.990	0.992	0.988	0.988	0.993
JUSTIFYING	0.980	0.976	0.977	0.974	0.984	0.985	0.984
MONOGLOSS	0.982	0.982	0.985	0.978	0.986	0.980	0.981
PROCLAIM	0.916	0.888	0.916	0.873	0.907	0.916	0.890
SOURCES	0.977	0.970	0.969	0.971	0.973	0.977	0.976

Note. Inter-item correlations are not simply the means of pairwise correlations in the Table. Inter-item correlation (rho) was calculated under the frequentist paradigm. Each pairwise Spearman correlation was computed under the Bayesian framework with a slightly skeptical prior distribution Beta(3, 3). This resulted in slightly conservative estimates of pairwise correlation coefficients.

4.3.4 Statistical analysis

All the main statistical analyses were conducted within the Bayesian framework (Gelman et al., 2015; McElreath, 2020). Unlike frequentist statistics, Bayesian statistical analysis has several conceptual and practical advantages (see also Nalborczyk et al., 2018; Winter & Bürkner, 2021).

From a conceptual standpoint, Bayesian inference is advantageous in its ability to quantify the uncertainties of parameters more intuitively (Dienes, 2011; McElreath, 2020). It has been observed that frequentist *p*-values and associated Confidence Intervals are often misused or misinterpreted as if they resulted from Bayesian inference (Dienes, 2011). In the frequentist paradigm, the population parameter value is considered "fixed" (i.e., there is a true parameter value), and Confidence Intervals quantify a range in which this "true" parameter can be captured at 95% (or any other predetermined percentage) of times given the infinite hypothetical resampling of data from the population (see Lambert, 2018, p. 130; McElreath, 2020, p. 58). For this reason, frequentist CIs do not allow a direct probabilistic interpretation (the value at the center of CIs is no more likely than another value at the edge of the range). On the other hand, Bayesian inference attempts to estimate the uncertainties of parameter values, given the available data and prior distributions. Thus, a posterior distribution allows the direct probabilistic interpretation of uncertainty of inference. For this reason, if one's interest lies in estimating parameters (not binary hypothesis testing), Bayesian inference can provide more relevant information about researchers' claims (Dienes, 2011). Because the goals of statistical analyses in second language research are often parameter estimations beyond null hypothesis testing (e.g., Norris, 2015; Plonsky & Oswald, 2014), Bayesian inference should be recognized as a more appropriate tool to achieve one's analytic purposes (e.g., Gudmestad et al., 2013; Norouzian et al., 2018, 2019).

From a practical point of view, a crucial advantage of the Bayesian approach is that the model makes it easier to converge with relatively complex random effects (Nalborczyk et al., 2018; Winter & Bürkner, 2021). Another practical advantage of the Bayesian approach to date includes a wide variety of response distributions that can be fitted with the help of a package

such as *brms* (Bürkner, 2017). This allows one to choose an appropriate response distribution family for one's analytical needs.

4.3.4.1 Model specifications

A series of the Poisson regressions were fitted to answer the research question, which concerns the relative distributions of Engagement strategies across disciplines, genres, and levels. The Poisson regression was deemed appropriate, considering the following characteristics of the underlying data-generating process: (a) frequency is expressed with positive values; (b) the upper bound frequency of Engagement strategies in a running text is not easily determined (there is no easily identified number of "trials"). Hox (2018) illustrates these characteristics using the number of typos in a book (Hox, 2018, p. 140), as we can talk about the differential rates at which typos occur within and across books. Given such characteristics, the Poisson regression is recommended to model occurrences of linguistic phenomena (Winter & Bürkner, 2021). The generic Poisson regression equation can be written as:

$$y_i = Poisson(\lambda_i)$$

$$ln(\lambda_i) = \alpha + \beta x_i$$

The generic formula for the Poisson regression indicates that, as with other families of GLM, we use a set of linear predictors (i.e., α and β, representing intercept and slopes, respectively) to predict a latent distribution of the outcome variable, or in the case of the Poisson, the event rate expressed with lambda (λ) (Heck & Thomas, 2020; Kruschke, 2014, Chapter 24; Winter & Bürkner, 2021). The canonical link function for Poisson regression is natural log (ln),

which allows positively skewed event rates (λ) predicted by a linear combination of predictors. Since log (1) equals 0, the predicted event rate will be 1 when the linear predictors α and β are fixed at zero. This, in turn, means that the exponent of the linear predictor will give us the predicted event rate (λ).

$$ln(\lambda_i) = \alpha + \beta x_i$$

$$\lambda_i = exp(\alpha + \beta x_i)$$

When modeling count data in language samples, the length of a particular sample (text length) must be controlled because it is intrinsically related to the number of occurrences of a linguistic phenomenon. In the register-oriented approach in corpus linguistics (e.g., Biber, Conrad, Reppen, et al., 2004), researchers have used normalized frequencies of a given construction (i.e., frequency of X per 1,000 words) to control for text length. Instead of using normed frequencies, I handled text lengths by including an offset term in the Poisson regression (Heck & Thomas, 2020; Winter & Bürkner, 2021). An offset term, also known as exposure, is a special type of covariate in the analysis, which attempts to control for the length of a particular sample in the analysis. This is done by fixing the regression coefficient of the offset term as a constant (i.e., 1), making the interpretations of other regression coefficients conditional on this offset term. In this study, I set the lengths of essays—calculated as the number of words divided by 1,000—as the offset term. This means that any predicted event rates of Engagement strategies are interpreted per 1,000 words. A normalizing constant of 1,000 words was chosen to follow the usual convention in corpus-based register analysis studies (e.g., Biber, Conrad, Reppen, et al., 2004).

In the present analysis, a regression model was formulated under a multilevel analysis of variance (ANOVA) approach (Gelman, 2005; Kruschke, 2014). In this approach, the effects of

any nominal variable (e.g., discipline, genre family) are estimated through so-called "random" effects. The multilevel (M)ANOVA approach has at least two important benefits for the present analysis. First, the multilevel MANOVA formulation allows the estimation of "sources" of variance and their higher-level interactions in a complex dataset. This is important in the analysis of the BAWE corpus because each factor had many levels for which to estimate variance (e.g., 13 genre families, 30 disciplines, and their interactions). In a fixed effects approach, the regression model specifies K – 1 pairs of comparison. As such, the resulting model does not offer omnibus summary statistics of variability due to a particular grouping of variables. In contrast, a multilevel (M)ANOVA approach allows the estimation of how much variability there is in the outcome due to combinations of grouping variables. Accordingly, the magnitude of random effects provides useful summary statistics of the variability around the grand mean (Gelman, 2005). Second, a multilevel formulation allows partial pooling of effects when estimating the marginal means of each group. This is particularly useful in the current context because the number of data points was not evenly distributed across combinations of genres and disciplines. In such a context, partially pooled estimates of marginal means for each group allow "borrowing" information from the entire distribution so that the estimates of small cells *shrink* towards the grand mean, thus counteracting potential extreme values due to the small sample size for each cell (Gelman & Hill, 2007). It is considered that partially pooled estimates thus result in more conservative and realistic estimates of specific group effects (Gelman & Hill, 2007; McElreath, 2020). Finally, relating to the previous points, a Bayesian multilevel formulation allows varying effects due to the grouping variable having a probability distribution (as a probability governing the distribution of "random" effects). For more details of the multilevel (M)ANOVA approach, see Gelman (2005).

4.3.4.2 Modeling procedure

Bayesian statistical modeling is a process of deriving posterior probability distributions of model parameters, given available data and prior information (Gelman, 2014; McElreath, 2020). To derive reliable posterior distributions, statisticians propose workflows for Bayesian analysis (Gelman et al., 2020; Schad, Betancourt, et al., 2021). The current analysis followed the general steps recommended in this literature, including prior predictive checks, model fit, model evaluation and validation, and model comparison.

4.3.4.2.1 Model construction and prior predictive checking

In any statistical inference, one must decide on a model to fit on available data. In the context of Generalized Linear Models (GLM; Agresti, 2015), one major decision in this model's construction stage is selecting a response distribution and which variables to include to capture the data-generating process. The next consideration at this stage in Bayesian inference is deciding on the prior distribution. Because the posterior distribution is sampled from a combination of data and prior distribution, the selection of the prior needs to be justified before models are fitted (Kruschke, 2014). Bayesian inference requires the researcher to set a prior distribution for every parameter to be estimated in the model. Several guidelines exist which help researchers reflect on their prior selections.

First, it is important to choose a prior that captures the possible ranges of model parameters and does not specify "absurd" ranges (McElreath, 2020). Taking a simple linear regression as an example, with appropriate domain knowledge, a researcher will expect the absolute value of a standard regression coefficient (beta) to rarely exceed ten, let alone 100. Thus, one can set a prior

distribution which makes it hard to produce such an extreme value (e.g., a normal distribution with a mean of zero and standard deviation of 5 will likely produce estimates over 10 for approximately 5% of a random draw from the prior).

This line of thinking must be applied to every parameter in the model, or at least every "class" of parameters, such as intercept, slope, sigma, and variance components. This process is called *prior predictive checking* (Gelman et al., 2020; McElreath, 2020). This step ensures that the selected prior will not lead to extreme parameter estimates. It is often the case that a combination of prior distributions may result in extreme predictions, given the entire parameter of the model. Thus, it is essential to check the implications of a specific prior choice for model prediction before fitting the data (Gelman et al., 2020; Schad, Betancourt, et al., 2021). The procedures recommended by Gelman et al. (2020) and Schad et al. (2021) were adopted for the current study.

When checking for the appropriacy of priors, it is also helpful to consider the levels of prior distributions. Gelman et al. (2021) list the possibilities of priors: "super-vague but proper prior; very weakly informative prior; generic weakly informative prior; specific informative prior" (p. 39). This means that one can justify the different levels of uncertainty expressed in a prior distribution depending on the goals and context of the research. For instance, it is possible and often recommended to use a set of "regularizing" priors to avoid extreme posterior estimates to counter-act small sample sizes. In this study, I used generic weakly informative priors to benefit from the regularizing effects of such priors (see Gelman, 2016 for recommended prior distributions for various research purposes). See section 4.4.2 for prior predictive checking.

4.3.4.2.2 Model fit and checking correct posterior approximation

In real-world data analysis, Bayesian posterior distributions are often difficult to derive analytically due to its highly multidimensional parameter space. For this reason, a range of approaches to approximate the posterior has been proposed, and one of the most current approaches includes simulation based on Markov Chain Monte Carlo (MCMC). A family of MCMCs is considered an efficient algorithm that tends to converge to stationary distributions in fewer steps (Gelman, 2014; McElreath, 2020).

In Bayesian inference, it is crucial to confirm whether the MCMC sampled has shown signs of convergence before being used for subsequent inferences. A few metrics allow researchers to carry out these checks. First, R-hat values quantify whether separately run MCMC chains stabilize into the same region (as an approximated global optimum). R-hat is calculated for each estimated parameter in the model, and often these values should be at most 1.05, but preferably below 1.01 (Vehtari et al., 2020). MCMC samples can also be examined through trace plots, where parameter values are plotted on the y-axis, and an MCMC iteration is plotted on the x-axis to show the transitions of the MCMC sampler over the iteration. Each fitted model must be checked against these criteria for valid Bayesian posterior estimations. If the MCMC chains do not converge, indicated by large R-hat values or diverging trace plots, this may indicate model specification errors that must be fixed before proceeding.

4.3.4.2.3 Model evaluation and validation—Posterior Predictive Checks and sensitivity analysis

After confirming the convergence of MCMC samples, one should examine whether the fitted model produces a reasonable prediction compared to the observed data. This process,

called Posterior Predictive Checks, ensures that the researcher evaluates whether their selection of various parts of the modeling strategy is adequate, including outcome distributions (i.e., Poisson over other variants of GLM distributions such as zero-inflated Poisson or negative binomial). When we detect a clear divergence of posterior distributions from the observed data, the statistical model must be revised (Gelman et al., 2020). In the current study, visual inspections of posterior predictive plots were conducted using bar graphs and Empirical Cumulative Distributional Function (ECDF) overlay graphs.

In addition to Posterior Predictive Checks, researchers need to check the influence of prior selection. In Bayesian analysis, after priors are narrowed down to theoretically permissible ranges, it is often recommended to test the impacts that different priors may have through sensitivity analysis (Gelman et al., 2020; Schad, Betancourt, et al., 2021). In this study, a few sets of prior distributions were compared regarding the ranges of random effects. This information is reported in the online supplementary materials (https://osf.io/dvyem/?view_only=7854a6e80f804740a3beac6fd36f6a17).

4.3.4.2.4 Interpretation

In a multilevel ANOVA, random effect parameters estimate the variability around the grand mean due to the grouping variable (in a standard deviation unit when the prior distribution is a normal distribution). Since the link function for Poisson regression is logarithmic, the parameter can be exponentiated to obtain the ratio of changes on the response scale (Gelman & Hill, 2007; Heck & Thomas, 2020). This means, for example, that when a slope parameter is 0.1, the corresponding change in the response event rate increases 10% (in multiplicative effects). In this study, the notion of Region Of Practical Equivalence (ROPE; Kruschke, 2014) was

employed to evaluate the practical significance of the regression parameters. Unlike the traditional null hypothesis testing (which tends to assess the *p*-values to see how much evidence there is to say the parameter is different from the null, often operationalized as zero), the ROPE approach allows the researchers to express the "null" hypothesis in terms of intervals, which may or may not including zero. The estimated intervals of parameter values are evaluated against these pre-set ROPE, not against a single value. One advantage of this approach includes the researchers can not only reject the null hypothesis, but also accept it or postpone the decisions regarding the null hypothesis. If the Credible Intervals (CrIs) exclude the pre-set ROPE, it indicates the rejection of the null hypothesis. A complete inclusion of the CrIs within the ROPE means that the parameter estimate is practically equivalent to the null hypothesis, thereby accepting it. A partial overlap means that the current data is not consistent with either, suggesting the inconclusive results. Although the appropriate range of ROPE values needs additional justification based the domain knowledge, it can help address some criticism against the null hypothesis testing—such as significant but practically negligible effect sizes. In other words, the use of ROPE allows one to test the hypothesis in terms of practically important effect size values (Kruschke, 2014; Vasishth & Gelman, 2021). In this study, the ROPE was specified on a log scale at 0.1, which approximately corresponds to 10% change in the predicted count (Kruschke, 2014; Makowski et al., 2019). This means that in order to declare the significant effect of a grouping variable, a standard deviation of the random effects must correspond to a variability in the occurrence rate by more than a 10%.

4.4 Results

4.4.1 Preliminary analysis

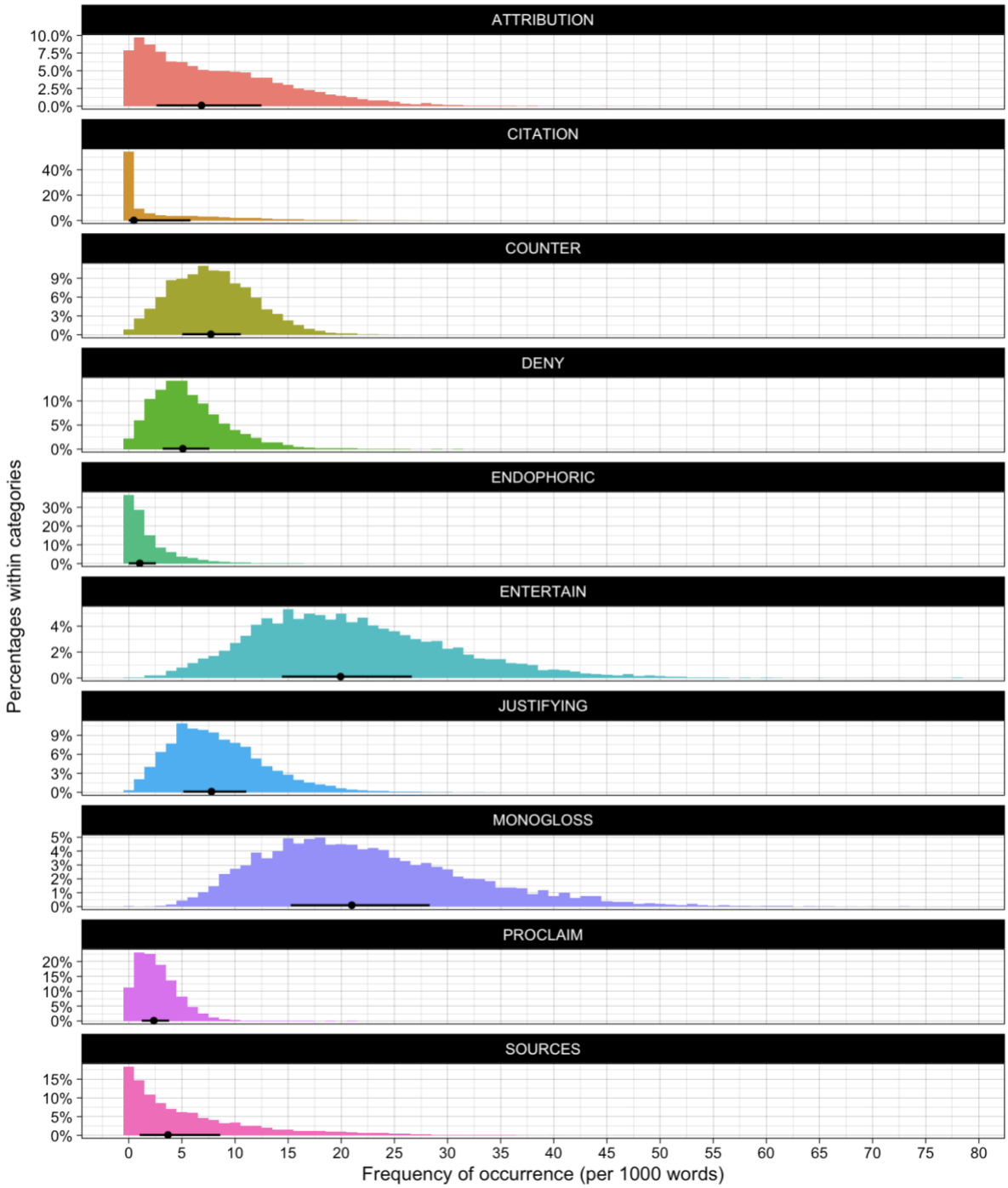
Before the main analyses, the counts for each Engagement strategy per document were examined (Fig. 4.1). This descriptive analysis served as information for prior specification in the next section. The normed frequency shown in Figure 4.1 (per 1,000 words) shows that it is rare for a writing assignment to contain a single Engagement strategy more frequently than 50 times per 1,000 words. The most frequently occurring category was MONOGLOSS, with a median of 21 times per 1,000 words. This was followed by ENTERTAIN (20 times), then COUNTER and JUSTIFYING. On the other hand, there were very infrequent CITATION, ENDOPHORIC, and PROCLAIM strategies with medians smaller than five times per 1,000 words.

4.4.2 Prior predictive checking

Next, a prior predictive check was conducted to ensure that the prior distribution captured the whole range of possible parameter ranges. From the descriptive plot in Figure 4.1, the predicted counts mostly fall within a range of 30 times per 1,000 words (in most categories) to 50 times at most (i.e., MONOGLOSS and ENTERTAIN). Thus, any predicted marginal means that exceeding these values are extremely unlikely. In other words, a permissive range of prior distributions will not produce very extreme values from these observed frequencies.

Figure 4.1

Descriptive plot of document-level frequencies of ten Engagement strategies (normed frequency per 1,000 words).

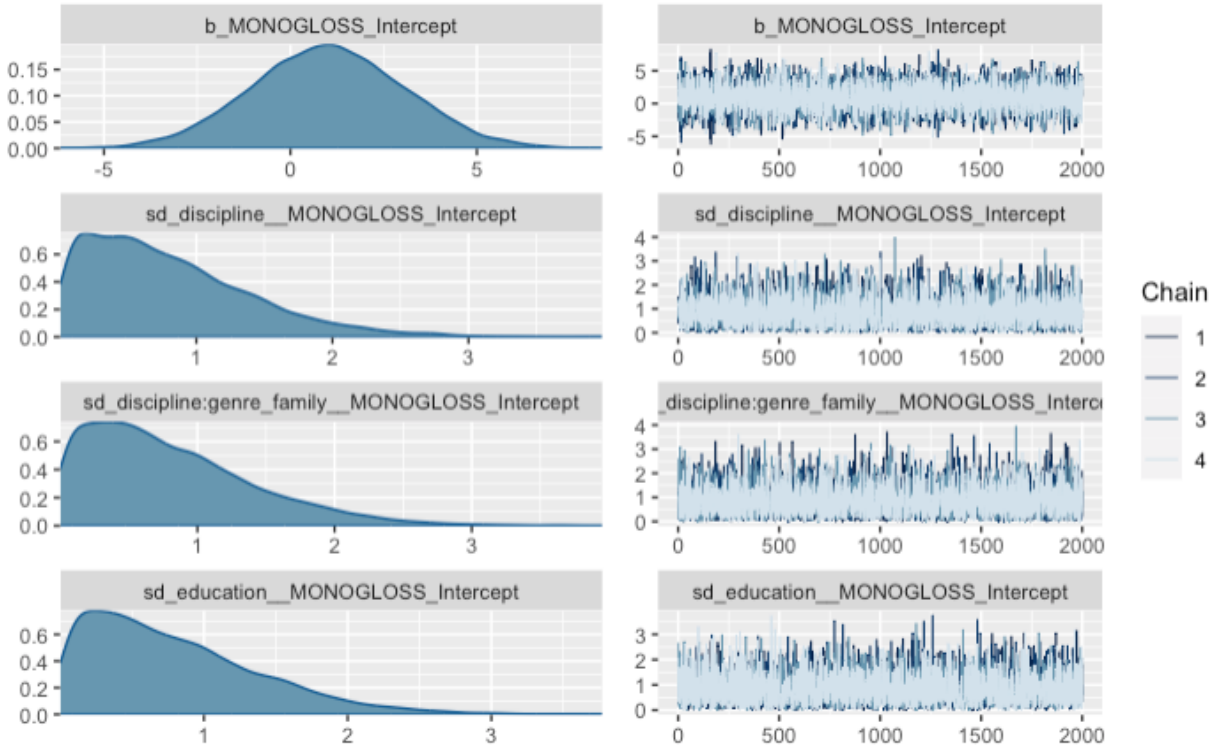


Note. This plot shows the document-level frequencies of each engagement strategy normalized to 1,000 words. The point in each panel indicates the median of the document-level normed frequency, the error bar shows the inter-quartile range.

As a first candidate prior, I tested a normal distribution with a mean of 1 and SD of 2 for any intercept parameter and a normal distribution with a mean of 0 and SD of 1. As we can see from Figure 4.2, this prior specification implies that the grand-mean estimates for each Engagement strategy will have a frequency of 2.718 with 95% interval estimates ranging from 0.0049 (exp[-3]) to 148.41 (= exp[5]). This covers a broader range than the descriptive plots in Figure 4.1, but at the same time, this does not produce theoretically implausible parameter estimates (such as a mean frequency of 2,000 times per 1,000 words). Combining these prior distributions of the grand-mean intercept with those of random effects, we can also simulate the implications of prior distributions for predicted counts (Fig. 4.3). It turned out that this prior specification would produce a permissible range of predictions with some regularizing effects around the grand means. In other words, it allows the estimates of means of each discipline to range from 0 to 40 times per 1,000 words, which is compatible with what is shown in the descriptive counts in Figure 4.1. For this reason, I proceeded with the main analysis using this prior distribution. Other candidate prior distributions, which were not used in the main analyses, can be found in the online supplementary material (https://osf.io/dvyem/?view_only=7854a6e80f804740a3beac6fd36f6a17).

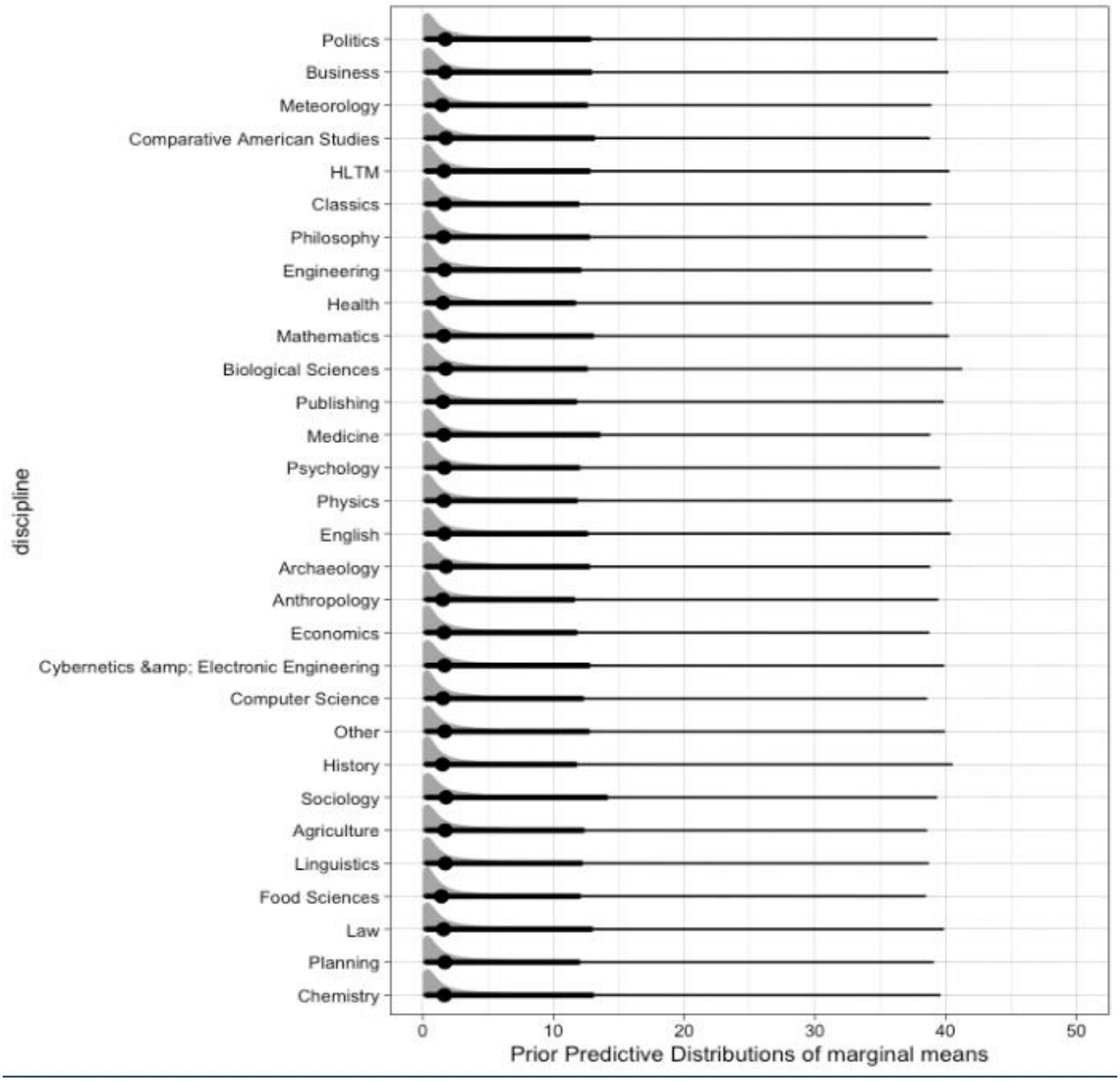
Figure 4.2

Prior predictive distribution of grand-mean frequency and random intercept estimates for the MONOGLOSS category.



Note. Prior distributions of four selected parameters are shown. The left panels show the density of the prior distributions; the right panels show trace plots, which are less relevant here. The current prior distribution suggests that the grand mean of the MONOGLOSS strategy rarely exceeds 148 times per 1,000 words (= exp(5)). The impact of grouping variables (random effects) rarely exceeds an SD of 2, which is equivalent to 7.389 times (= exp(2)).

Figure 4.3
Prior predictive distributions of marginal means of MONOGLOSS frequency by discipline.



4.4.3 Multilevel MANOVA

A multilevel MANOVA model was fitted using the prior distribution shown above. All R-hat values suggest MCMC convergence, as they are below a conservative threshold of 1.01 (Vehtari et al., 2020). Table 4.5 presents the results for the multilevel MANOVA model. Figure 4.4 presents these variance estimates as ratios of event rates and 50% (thick lines) and 95%

Credible Intervals (thin lines). The gray areas in these figures indicate the ROPE, defined as less than 1.1 on a logged event rate scale. According to Table 4.5, the grand-mean event rates were highest for ENTERTAIN (3.035; 95% CrIs = [2.547; 3.443]) and MONOGLOSS (3.000; 95% CrI = [2.621; 3.312]). The predicted counts correspond to 20.80 ($e^{3.035}$) and 20.08 ($e^{3.000}$) times per 1000 words. The least frequent category was CITATION, whose grand-mean event rate was only 0.063, or 1.065 times per 1,000 words.

Group-level effects showed relatively large effects of writers and two-way interactions between genre family and disciplines for Engagement strategies (excluding the ROPE in all the 95% CrIs). In particular, it is noteworthy that two-way interactions tended to be larger than the main effects of disciplines or genre families in many cases (see Fig. 4.4 for a visualization) and sometimes as large as the variation due to individual writers. This suggests that combinations of these two contextual factors may be more important in writers' selection of Engagement strategies than independent effects of disciplines or genres alone. In contrast, the effects of grades, levels, and writers' secondary education were minor, and their estimates were associated with large uncertainties (likely due to the small number of levels). The writer's first language (L1) affected three categories (DENY, JUSTIFYING, CITATION).

Table 4.5
Summary of multilevel MANOVA.

Intercepts		*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
	MONOGLOSS	3.000	0.173	2.621	3.312	1.000	6391	4402
	ATTRIBUTION	1.485	0.277	0.931	2.019	1.001	5239	4233
	ENTERTAIN	3.035	0.225	2.547	3.443	1.000	6581	3954
	PROCLAIM	0.666	0.280	0.093	1.236	1.000	7533	5531
	DENY	1.604	0.274	1.041	2.162	1.001	6962	4895
	COUNTER	1.836	0.141	1.588	2.089	1.000	5710	4156
	JUSTIFYING	1.982	0.222	1.539	2.383	1.001	6645	4782
	CITATION	0.063	0.860	-1.604	1.800	1.001	7463	5452
	ENDOPHORIC	0.175	0.333	-0.455	0.873	1.000	7381	5481
	SOURCES	0.837	0.298	0.245	1.419	1.002	5389	5197
Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~discipline (# of levels: 32)								
	MONOGLOSS	0.142	0.031	0.087	0.209	1.001	3121	3765
	ATTRIBUTION	0.377	0.080	**0.236**	**0.549**	1.002	3429	5254
	ENTERTAIN	0.202	0.044	**0.124**	**0.298**	1.001	3079	4750
	PROCLAIM	0.181	0.066	0.048	0.311	1.007	1374	2012
	DENY	0.184	0.050	0.090	0.286	1.001	2037	2243
	COUNTER	0.123	0.046	0.026	0.212	1.001	1458	1565
	JUSTIFYING	0.172	0.051	0.071	0.274	1.002	1640	1923
	CITATION	1.395	0.204	**1.048**	**1.845**	1.001	5529	5465
	ENDOPHORIC	0.462	0.112	**0.262**	**0.696**	1.001	2252	3046
	SOURCES	0.522	0.104	**0.341**	**0.752**	1.000	3215	4939
Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~genre_family (Number of levels: 13)								
	MONOGLOSS	0.150	0.045	0.085	0.257	1.000	4369	5144
	ATTRIBUTION	0.538	0.136	**0.335**	**0.855**	1.002	5405	5712
	ENTERTAIN	0.201	0.057	**0.115**	**0.334**	1.000	5520	5043
	PROCLAIM	0.234	0.072	**0.123**	**0.404**	1.001	3954	4511
	DENY	0.233	0.068	**0.132**	**0.396**	1.000	5298	5805
	COUNTER	0.242	0.062	**0.149**	**0.389**	1.000	4913	4553
	JUSTIFYING	0.169	0.052	0.087	0.288	1.000	4642	5342
	CITATION	0.809	0.206	**0.491**	**1.293**	1.001	5529	4972
	ENDOPHORIC	0.450	0.132	**0.241**	**0.753**	1.001	4393	5332
	SOURCES	0.625	0.156	**0.386**	**1.004**	1.001	4990	5335

Table 4.5 (Cont'd)

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~discipline:genre_family (# of levels: 228)								
	MONOGLOSS	0.223	0.013	**0.198**	**0.250**	1.000	3778	5403
	ATTRIBUTION	0.619	0.039	**0.549**	**0.698**	1.002	3516	5247
	ENTERTAIN	0.301	0.018	**0.267**	**0.339**	1.000	3416	5308
	PROCLAIM	0.470	0.033	**0.407**	**0.540**	1.000	3470	5316
	DENY	0.383	0.025	**0.336**	**0.437**	1.001	3155	4967
	COUNTER	0.283	0.021	**0.246**	**0.327**	1.003	2877	4681
	JUSTIFYING	0.304	0.021	**0.265**	**0.348**	1.001	3100	5080
	CITATION	0.993	0.073	**0.862**	**1.149**	1.000	3395	4923
	ENDOPHORIC	0.843	0.059	**0.735**	**0.966**	1.002	2906	4724
	SOURCES	0.762	0.050	**0.670**	**0.865**	1.002	3413	4578

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~grade (Number of levels: 3)								
	MONOGLOSS	0.131	0.183	0.004	0.657	1.004	2125	4345
	ATTRIBUTION	0.233	0.246	0.034	0.937	1.000	4451	5203
	ENTERTAIN	0.090	0.141	0.010	0.503	1.001	4843	5292
	PROCLAIM	0.195	0.226	0.018	0.863	1.000	4073	5446
	DENY	0.296	0.276	0.053	1.074	1.000	4182	5188
	COUNTER	0.083	0.154	0.003	0.509	1.001	2959	3616
	JUSTIFYING	0.129	0.185	0.013	0.654	1.001	4465	5725
	CITATION	1.245	0.435	**0.614**	**2.282**	1.001	6659	5516
	ENDOPHORIC	0.190	0.226	0.021	0.866	1.000	4574	6529
	SOURCES	0.198	0.236	0.026	0.903	1.000	4458	5869

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~level (Number of levels: 4)								
	MONOGLOSS	0.108	0.101	0.034	0.366	1.000	5259	5466
	ATTRIBUTION	0.041	0.051	0.010	0.150	1.000	6498	5818
	ENTERTAIN	0.298	0.204	**0.107**	**0.864**	1.001	3830	4203
	PROCLAIM	0.162	0.139	0.046	0.541	1.001	4936	5964
	DENY	0.176	0.138	0.055	0.554	1.000	5452	5324
	COUNTER	0.031	0.044	0.001	0.132	1.000	2706	3576
	JUSTIFYING	0.196	0.159	0.064	0.638	1.001	4975	4839
	CITATION	0.481	0.264	**0.194**	**1.183**	1.001	7723	5682
	ENDOPHORIC	0.324	0.211	**0.116**	**0.919**	1.000	5876	6105
	SOURCES	0.028	0.038	0.003	0.110	1.000	4903	3947

Table 4.5 (Cont'd)

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~L1 (Number of levels: 42)								
	MONOGLOSS	0.137	0.033	0.082	0.213	1.000	2522	3756
	ATTRIBUTION	0.163	0.065	0.049	0.306	1.007	951	1716
	ENTERTAIN	0.125	0.028	0.079	0.188	1.001	2656	4078
	PROCLAIM	0.149	0.045	0.072	0.252	1.001	2126	3992
	DENY	0.233	0.053	**0.142**	**0.350**	1.000	2616	5075
	COUNTER	0.098	0.032	0.044	0.170	1.004	1471	2988
	JUSTIFYING	0.192	0.040	**0.123**	**0.281**	1.001	2504	4180
	CITATION	0.800	0.197	**0.468**	**1.241**	1.002	2074	4409
	ENDOPHORIC	0.139	0.079	0.010	0.313	1.007	700	2006
	SOURCES	0.168	0.066	0.064	0.320	1.003	1068	2401

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~Secondary Education (Number of levels: 14)								
	MONOGLOSS	0.076	0.052	0.006	0.207	1.003	1790	2490
	ATTRIBUTION	0.111	0.105	0.003	0.393	1.001	1676	3041
	ENTERTAIN	0.128	0.056	0.047	0.264	1.000	3113	4159
	PROCLAIM	0.142	0.111	0.006	0.410	1.002	2202	4254
	DENY	0.165	0.100	0.027	0.402	1.000	2171	2645
	COUNTER	0.050	0.050	0.002	0.182	1.000	2014	3194
	JUSTIFYING	0.108	0.059	0.026	0.253	1.002	2466	2746
	CITATION	0.532	0.313	0.062	1.273	1.002	1563	3427
	ENDOPHORIC	0.151	0.135	0.005	0.504	1.002	1797	3037
	SOURCES	0.163	0.150	0.006	0.562	1.003	1330	2783

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~student_id (Number of levels: 620)								
	MONOGLOSS	0.283	0.009	**0.266**	**0.301**	1.002	2484	4230
	ATTRIBUTION	0.541	0.020	**0.504**	**0.580**	1.001	1700	3388
	ENTERTAIN	0.303	0.010	**0.285**	**0.323**	1.001	2066	4014
	PROCLAIM	0.453	0.016	**0.421**	**0.485**	1.001	2527	3856
	DENY	0.359	0.013	**0.335**	**0.385**	1.001	2465	4222
	COUNTER	0.284	0.010	**0.265**	**0.304**	1.001	2646	4748
	JUSTIFYING	0.391	0.013	**0.367**	**0.417**	1.003	2227	2986
	CITATION	1.879	0.068	**1.750**	**2.020**	1.001	1993	3365
	ENDOPHORIC	0.770	0.028	**0.717**	**0.825**	1.001	2461	4537
	SOURCES	0.696	0.025	**0.649**	**0.746**	1.004	2093	4002

Table 4.5 (Cont'd)

Factor	Category	*Estimate*	*Est.Error*	*LL*	*UL*	*R-hat*	*Bulk ESS*	*Tail ESS*
~model (Number of levels: 4)								
	MONOGLOSS	0.116	0.109	0.037	0.391	1.001	2248	3797
	ATTRIBUTION	0.099	0.099	0.031	0.353	1.000	2397	3597
	ENTERTAIN	0.052	0.056	0.016	0.184	1.001	2209	3825
	PROCLAIM	0.271	0.188	0.097	0.813	1.001	2904	4395
	DENY	0.075	0.077	0.023	0.262	1.002	2346	3209
	COUNTER	0.051	0.056	0.015	0.180	1.001	2364	3377
	JUSTIFYING	0.141	0.133	0.046	0.467	1.001	2401	3787
	CITATION	0.179	0.149	0.059	0.570	1.001	2555	4266
	ENDOPHORIC	0.110	0.102	0.035	0.361	1.001	2003	3351
	SOURCES	0.084	0.081	0.026	0.295	1.002	2043	3607

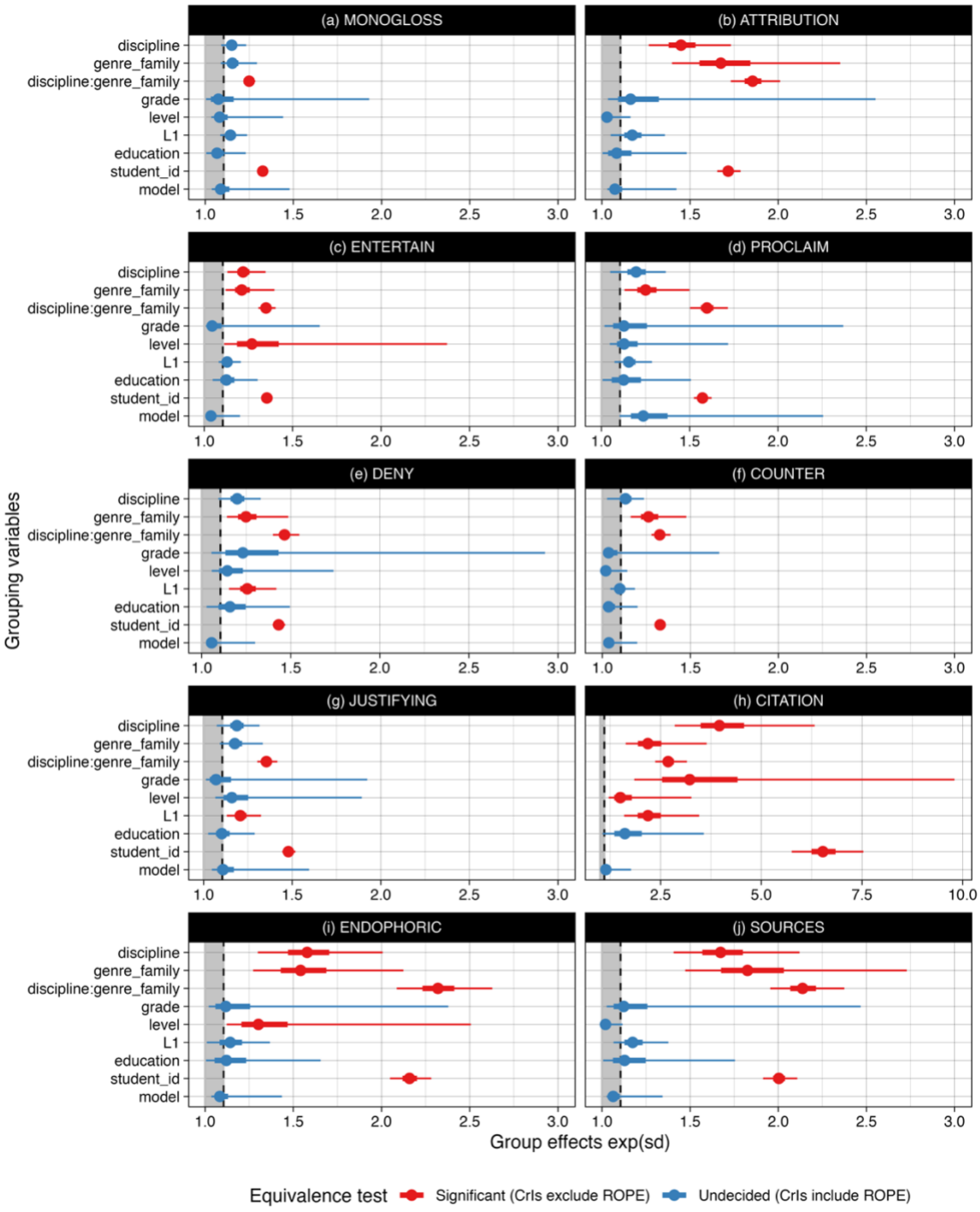
Note. LL = Lower limit of 95% Credible Intervals; UL = Upper limit of 95% Credible Intervals; boldface letters indicate CrIs that did not overlap the Region of Practical Equivalence (set at 0.1).

4.4.3.1 MONOGLOSS

Figure 4.4 (panel a) compares group effects on the MONOGLOSS strategy. Based on the ROPE, two effects are found to be significant as 95% CrIs exclude it. These two are discipline * genre family interaction (Median = 0.223; 95% CrIs = [0.198; 0.250]), and writer (Median = 0.283; 95% CrIs = [0.266; 0.301]). The interaction effect indicates that the variability in the frequencies of MONOGLOSS around the grand mean due to the discipline-by-genre family combination is estimated to be 1.25 times (CrIs = [1.218; 1.284]) a Standard Deviation unit. Similarly, the main effect of writer indicates that the variability of MONOGLOSS due to individual writers is estimated to range from 1.304 to 1.351 times with a median of 1.327. Other factors include the ROPE in their CrIs and do not fully overlap it, indicating that more evidence is needed to reject or accept the null hypothesis (Kruschke, 2014).

Figure 4.4

Group-level effects (random intercepts) in exponential scale



Note. The graph shows posterior estimates of the magnitude of grouping factors in the distribution of each engagement strategy. The dots indicate the posterior median; thick lines show 50% CrIs, and thin lines show 95% CrIs. Gray areas indicate the Region of Practical Equivalence (ROPE). When 95% CrIs did not include the ROPE, the current analysis concluded that there is significant variability in the outcome due to the grouping factor (Kruschke, 2014).

4.4.3.2 ATTRIBUTION

Panel (b) of Figure 4.4 compares the variabilities due to each grouping variable and their estimated uncertainties for the ATTRIBUTION strategy. Four effects are significant as they exclude the ROPE. These are discipline (Median = 0.377; 95% CrI = [0.236; 0.549]), genre family (Median = 0.538; 95% CrI = [0.335; 0.855]), discipline * genre family interaction (Median = 0.619; 95% CrI = [0.549; 0.698]), and writer (Median = 0.541; 95% CrI = [0.504; 0.580]). Other factors include the ROPE in their CrIs, which do not fully overlap it, indicating that there is not enough evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.3 ENTERTAIN

According to panel (c) of Figure 4.4, five sources of variance are significant, excluding the ROPE in their 95% CrIs. These are discipline (Median = 0.202; 95% CrI = [0.124; 0.298]), genre family (Median = 0.201; 95% CrI = [0.115; 0.334]), discipline * genre family interaction (Median = 0.301; 95% CrI = [0.267; 0.339]), course level (Median = 0.298; 95% CrI = [0.107; 0.864]), and writer (Median = 0.303; 95% CrI = [0.285; 0.323]). Other factors include the ROPE in their CrIs and do not fully overlap it, indicating that there is not enough evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.4 PROCLAIM

Three factors are significant, excluding the ROPE in their 95% CrIs (Panel (d) of Figure 4). These are genre family (Median = 0.234; 95% CrI = [0.123; 0.404]), discipline * genre family interaction (Median = 0.470; 95% CrI = [0.407; 0.540]), and writer (Median = 0.453; 95% CrI = [0.421; 0.485]). Other factors include the ROPE in their CrIs, and do not fully overlap it,

indicating that more evidence is needed to reject or accept the null hypothesis (Kruschke, 2014). It is also noteworthy that the effect of the Engagement Analyzer model is approaching significance as it only hits the ROPE at the edge of their CrIs (where the posterior probability is quite low). This suggests that measurement error may be as large as other effects.

4.4.3.5 DENY

Panel (e) of Figure 4.4 compares variabilities due to each grouping variable for the DENY strategy. Four factors are found to be significant as they do not include the ROPE in their 95% CrIs. These are genre family (Median = 0.233; 95% CrI = [0.132; 0.396]), discipline * genre family interaction (Median = 0.383; 95% CrI = [0.336; 0.437]), writer's L1, and writer (Median = 0.233; 95% CrI = [0.142; 0.350]). Other factors include the ROPE in their CrIs, indicating there is insufficient evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.6 COUNTER

Three factors are found to be significant as they exclude the ROPE from their 95% CrIs. These are genre family (Median = 0.242; 95% CrI = [0.149; 0.389]), discipline * genre family interaction (Median = 0.283; 95% CrI = [0.246; 0.327]), and writer (Median = 0.284; 95% CrI = [0.265; 0.304]). Other factors include the ROPE in their CrIs, indicating there is insufficient evidence to reject or accept the null hypothesis (Kruschke, 2014). Most of the 95% CrI of level was inside the ROPE, indicating that with replication and more data, the result may be in line with the null hypothesis.

4.4.3.7 JUSTIFYING

Three significant sources of variance are identified for the JUSTIFYING strategy—discipline * genre family interaction (Median = 0.304; 95% CrI = [0.265; 0.348]), writer's L1 (Median = 0.192; 95% CrI = [0.123; 0.281]), and writer (Median = 0.391; 95% CrI = [0.367; 0.417]). Other factors include the ROPE in their CrIs, indicating there is insufficient evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.8 CITATION

A total of seven grouping variables are found to be significant—discipline (Median = 1.395; 95% CrI = [1.048; 1.845]), genre family (Median = 0.809; 95% CrI = [0.491; 1.293]), discipline * genre family interaction (Median = 0.993; 95% CrI = [0.862; 1.149]), grades (Median = 1.245; 95% CrI = [0.614; 2.282]), level (Median = 0.481; 95% CrI = [0.194; 1.183]), L1 (Median = 0.800; 95% CrI = [0.468; 1.241]), and writer (Median = 0.800; 95% CrI = [0.468; 1.241]). Other factors include the ROPE in their CrIs, indicating that there is not enough evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.9 ENDOPHORIC

Five sources of variance are significant for ENDOPHORIC strategies—discipline (Median = 0.462; 95% CrI = [0.262; 0.696]), genre family (Median = 0.450; 95% CrI = [0.241; 0.753]), discipline * genre family interaction (Median = 0.843; 95% CrI = [0.735; 0.966]), course level (Median = 0.324; the 95% CrI = [0.116; 0.919]), and writer (Median = 0.770; the 95% CrI = [0.717; 0.825]). Other factors include the ROPE in their CrIs, indicating that there is not enough evidence to reject or accept the null hypothesis (Kruschke, 2014).

4.4.3.10 SOURCES

Four sources of variance are significant for the SOURCES strategy—discipline (Median = 0.522; 95% CrI = [0.341; 0.752]), genre family (Median = 0.625; 95% CrI = [0.386; 1.004]), discipline * genre family interaction (Median = 0.762; 95% CrI = [0.670; 0.865]), and writer (Median = 0.696; 95% CrI = [0.649; 0.746]). Other factors include the ROPE in their CrIs, indicating that there is not enough evidence to reject or accept the null hypothesis (Kruschke, 2014). Also, most of the 95% CrIs for course level fell within the ROPE, indicating a replication with more data may find a zero-effect due to the course level.

4.5 Discussion

4.5.1 Summary of findings

The multilevel MANOVA indicates that two sources of variance tend to account for major parts of the distributions of Engagement strategies in university writing assignments—the interaction between genre family and disciplines and individual writers (Fig. 4.4). In most categories of Engagement, two-way interaction tends to explain more considerable variance than the main effect. In subsets of Engagement strategies, such as MONOGLOSS and JUSTIFYING, the main effects of disciplines and genre families are NOT significantly different from the ROPE. These more significant effects of interaction compared to the main effects suggest that specific combinations of disciplines and genre families are a strong factor motivating the selection of Engagement strategies. On the other hand, the use of Engagement strategies varies to a lesser degree due to other sources of variance, such as course level, assignment grade, the writer's secondary education, and the writer's L1. The remainder of this chapter is organized as follows. First, possible reasons for the sizeable two-way interaction effect of discipline and genre

family are discussed. The second part of the discussion focuses on the non-significant effects of assignment-related and individual-related factors, particularly in light of the large individual variabilities due to writers. The chapter then concludes with implications for English for Academic Purposes (EAP) courses and suggestions for future research.

4.5.2 Discipline and Genre Family Combinations to Explain Registers of University Written Assignments

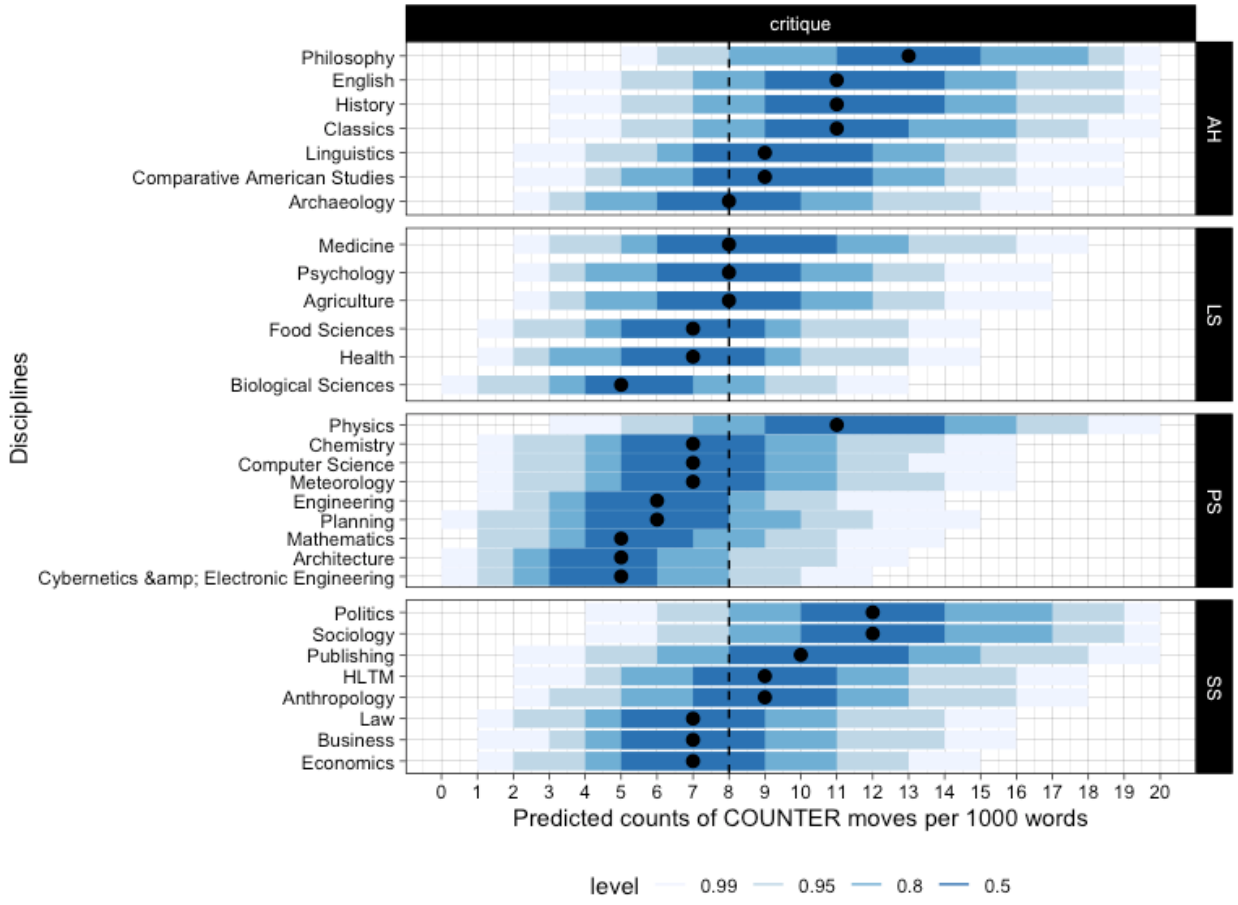
The significant interaction between genre family and discipline can be interpreted in several ways. First, the findings indicate that university writing assignments are far more nuanced and complex than simple characterizations based on genres or disciplines. From a genre-based pedagogy perspective, there may be several possible approaches to writing a single genre, which is likely influenced by several specific contextual factors, including disciplinary-specific communicative needs. To illustrate this idea of disciplinary-specific communication needs, Figure 4.5 visualizes the predicted frequencies of COUNTER strategies by a hypothetical average person from each discipline for the critique genre family (the uncertainties of marginal means were much tighter). As can be seen, there is considerable variation in the predicted frequency of COUNTER strategies across disciplines, even within the same genre family. For example, COUNTER strategies tend to be more frequent in Philosophy, Politics, Sociology, and other Social Sciences and Arts and Humanity disciplinary groups, but less frequent in Electric Engineering, Architecture, and Biological Sciences. Similarly, within-genre family variability can be illustrated for ENTERTAIN in case studies (Fig. 4.6). This shows that ENTERTAIN is more frequent in case studies in Economics, Sociology, and Health. In contrast, this strategy is less frequent in case studies from Food sciences, Computer science, and Law (the 99th

percentiles in the latter disciplines are less frequent than the medians of most frequent disciplines). Overall, in a genre-based approach, the findings suggest that there is variation within a genre in terms of the use of Engagement strategies across disciplines.

From a disciplinary writing perspective, the same interaction may speak to potential variations within disciplinary writing, although there appears to be less variability in this view. For example, the frequencies of ATTRIBUTION vary to a certain extent due to the genre families within Psychology writing (see Fig. 4.7). In psychology, proposal and empathy writing appear to elicit ATTRIBUTION more frequently than other genres, such as critique, explanation, and methodology recounts. However, in Agriculture writing (Figure 4.8), the ranking of genre families differs from that of psychology. For example, some genres that elicit frequent ATTRIBUTIONS (e.g., empathy writing, critique) for psychology tend to rank lower for Agriculture. Conversely, genres with relatively few ATTRIBUTIONS in psychology writing (e.g., exercise) rank higher in Agriculture.

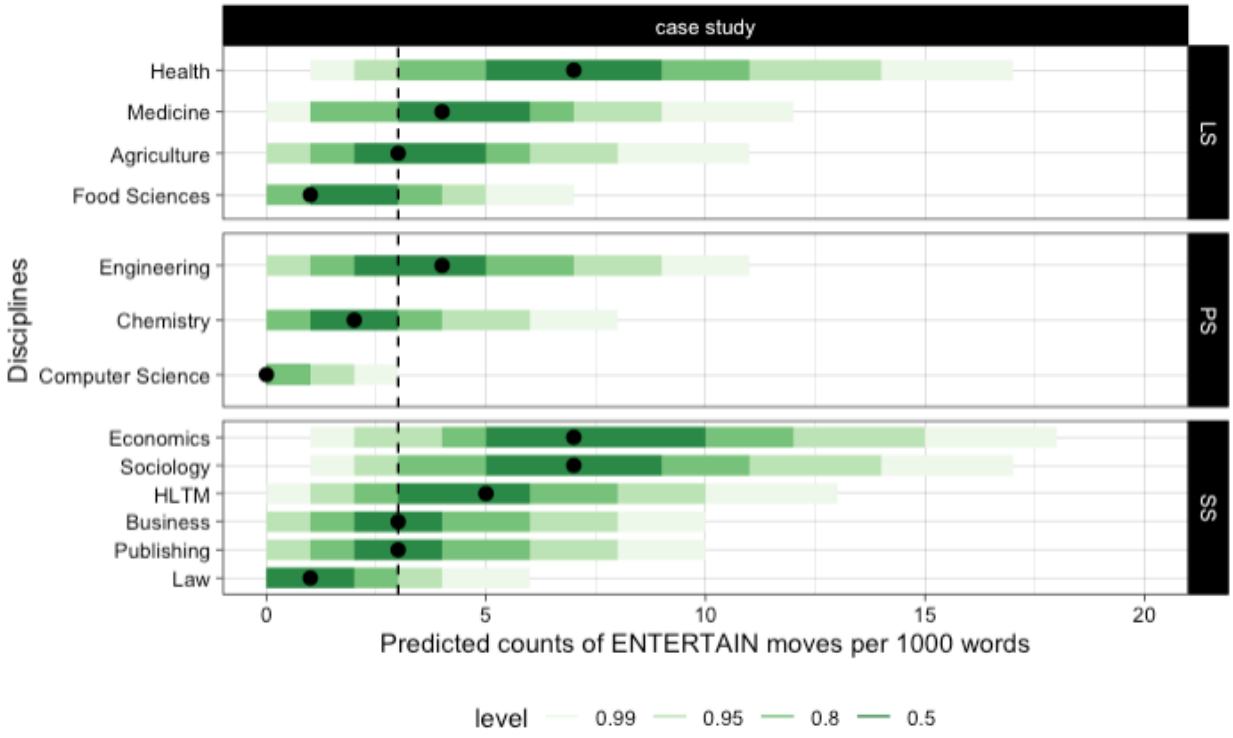
In summary, the current findings indicate that the distribution of Engagement strategies tends to vary not only due to genres or disciplines independently, but also due to combinations of these two factors. This means that the notion of "genres" may be nuanced, and registers may be chaptered by more precise definitions of genre and discipline combinations. See online supplementary materials for the remaining prediction plots; https://osf.io/dvyem/?view_only=7854a6e80f804740a3beac6fd36f6a17). More specific implications of these findings are addressed in the implications section.

Figure 4.5
Predicted counts of COUNTER strategies in the critique genre family.



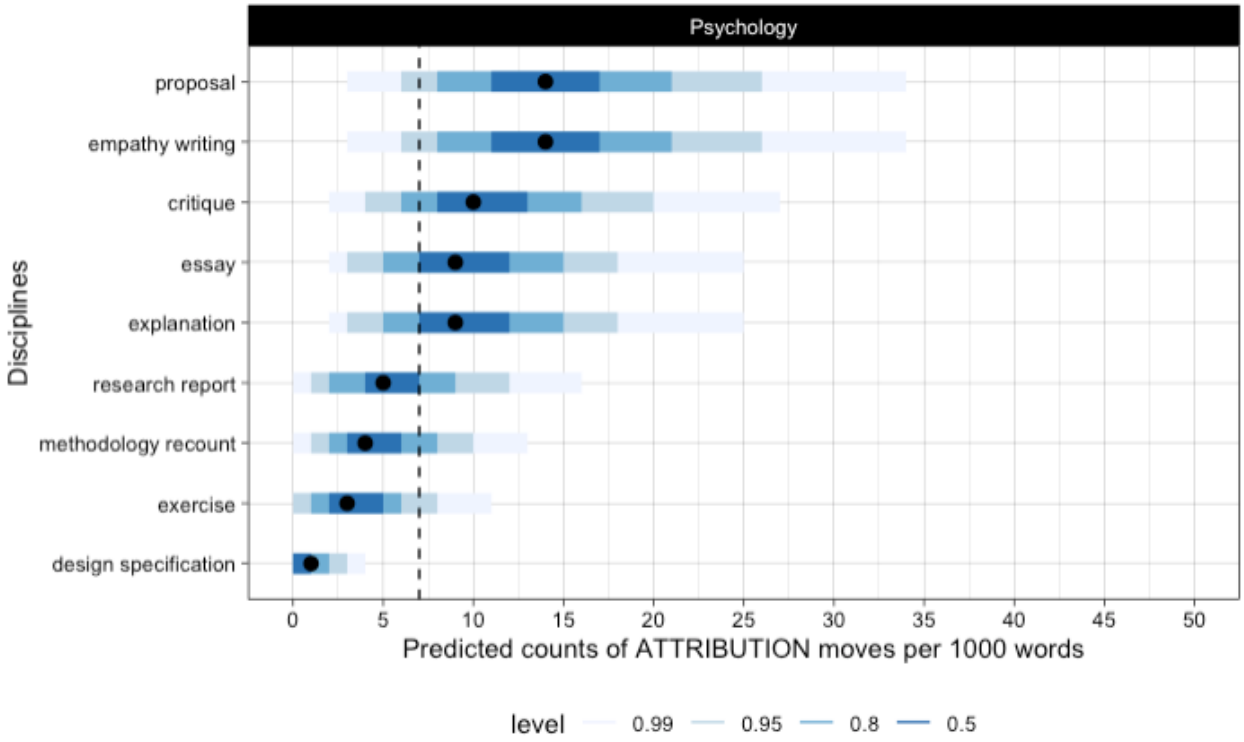
Note. Prediction intervals based on the multilevel MANOVA model are plotted. Model-based predictions were generated with two-way interactions of genre family and disciplines, and the main effects of these two while other random effects were integrated out from the original formula. The vertical dotted line shows the overall median across the disciplines plotted. The dots indicate the median of the model-based prediction (the most probable frequency), and graded intervals show the ranges in which model-based prediction values fall along with respective probabilities (50%, 80%, 95%, 99%). For example, a predicted assignment in the 50th percentile for Biological Sciences uses COUNTER strategies 5 times per 1,000 words, which corresponds to the 5th percentile in English, History, and Classics writing. Note that prediction intervals are typically wider than Credible Intervals because the latter only show the uncertainty of the central tendency.

Figure 4.6
Predicted counts of ENTERTAIN in the case study genre family.



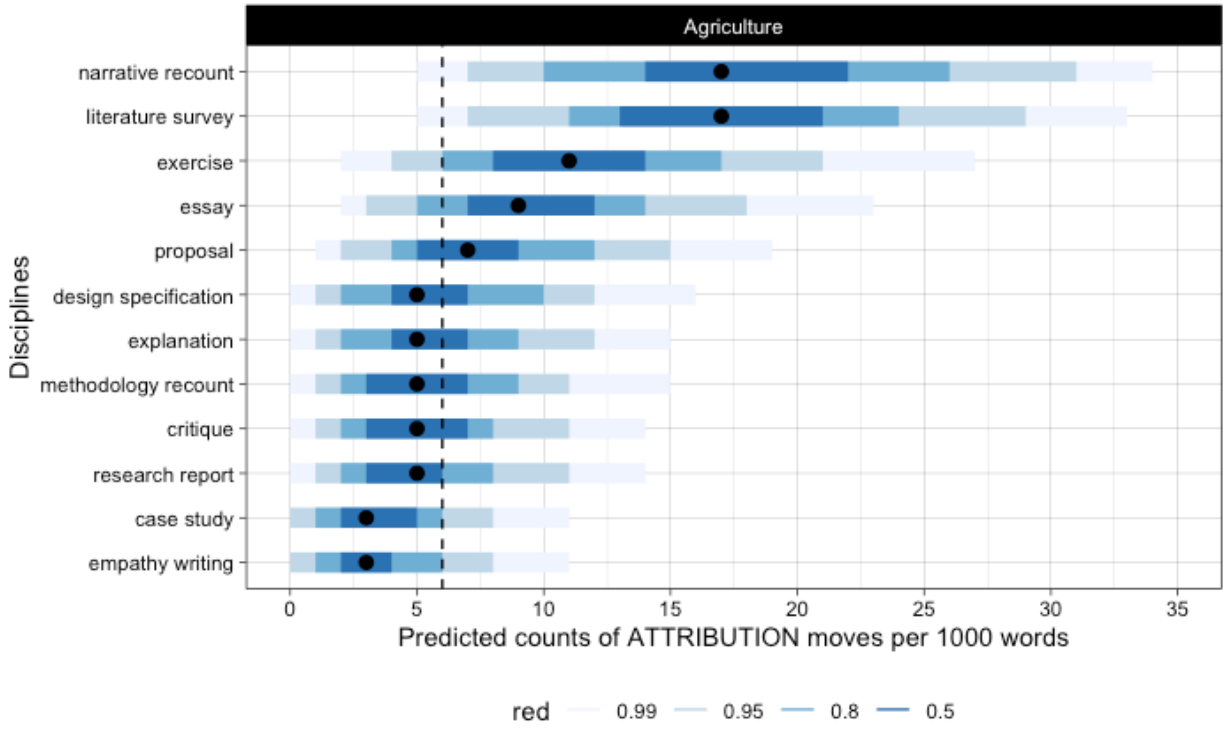
Note. Prediction intervals based on the multilevel MANOVA model are plotted. As in Figure 4.5, model-based predictions were generated with two-way interactions of genre family and disciplines, and the main effects of these two, while other random effects were integrated out from the original formula. The vertical dotted line shows the overall median across the disciplines plotted. The dots indicate the median of the model-based prediction (the most probable frequency), and graded intervals show the ranges in which model-based prediction values fall along with respective probabilities (50%, 80%, 95%, 99%). Note that prediction intervals are typically wider than Credible Intervals because the latter only show the uncertainty of the central tendency.

Figure 4.7
Predicted counts of ATTRIBUTION within Psychology writing.



Note. Prediction intervals based on the multilevel MANOVA model are plotted. As in Figure 5, model-based predictions were generated with two-way interactions of genre family and disciplines, and the main effects of these two, while other random effects were integrated out from the original formula. The vertical dotted line shows the overall median across the genre families plotted. The dots indicate the median of the model-based prediction (the most probable frequency), and graded intervals show the ranges in which the model-based prediction values fall along with respective probabilities (50%, 80%, 95%, 99%). Note that prediction intervals are typically wider than Credible Intervals because the latter only show the uncertainty of the central tendency.

Figure 4.8
Predicted counts of ATTRIBUTION within Agriculture writing.



Note. Prediction intervals based on the multilevel MANOVA model are plotted. As in Figure 5, model-based predictions were generated with two-way interactions of genre family and disciplines, and the main effects of these two, while other random effects were integrated out from the original formula. The vertical dotted line shows the overall median across the genre families plotted. The dots indicate the median of the model-based prediction (the most probable frequency), and graded intervals show the ranges in which model-based prediction values fall along with respective probabilities (50%, 80%, 95%, 99%). Note that prediction intervals are typically wider than Credible Intervals because the latter only show the uncertainty of the central tendency.

4.5.3 Possible Accounts for the Smaller Impact of Writer-related Factors—L1 and Education

The multivariate MANOVA also indicates that assignment grades, course levels, writers' previous education, and their L1s are not among the most influential factors differentiating Engagement strategies compared to individual writers. Although the fact that individual writers explain more variance than any other higher groupings, such as L1s and their secondary education system, is not surprising, the amounts of variability due to these writer-related factors is relatively small. For example, the effects of the writer's L1 exceed the ROPE only in DENY, JUSTIFYING, and CITATION, and no strategies differ greatly due to their secondary education. This suggests that, in the current corpus, contextual parameters more directly relevant to the communicative purposes of writing are more important. This is likely because the distribution of Engagement strategies primarily concerns the choice of discourse meaning, rather than the linguistic forms to achieve these communicative discourse functions. In one explanation, we can argue that the sheer counts of these strategies may not explain the nuanced ways in which L1 influence surfaces. Several possible ways in which the writer's L1 and their educational background matter are listed in directions for future research.

4.5.4 Engagement as Writing Style or Writing Quality?

Beyond higher-level writer-related factors, one reason why individual writers may explain more variance than any other higher-level factors is writers' idiosyncratic styles or originality in their argumentative patterns. In corpus-based register analysis, researchers refer to these individual variations as "styles" because their patterning seems to go beyond the effects of other contextual variables (e.g., genre, discipline, grades, or course levels). In other words, the

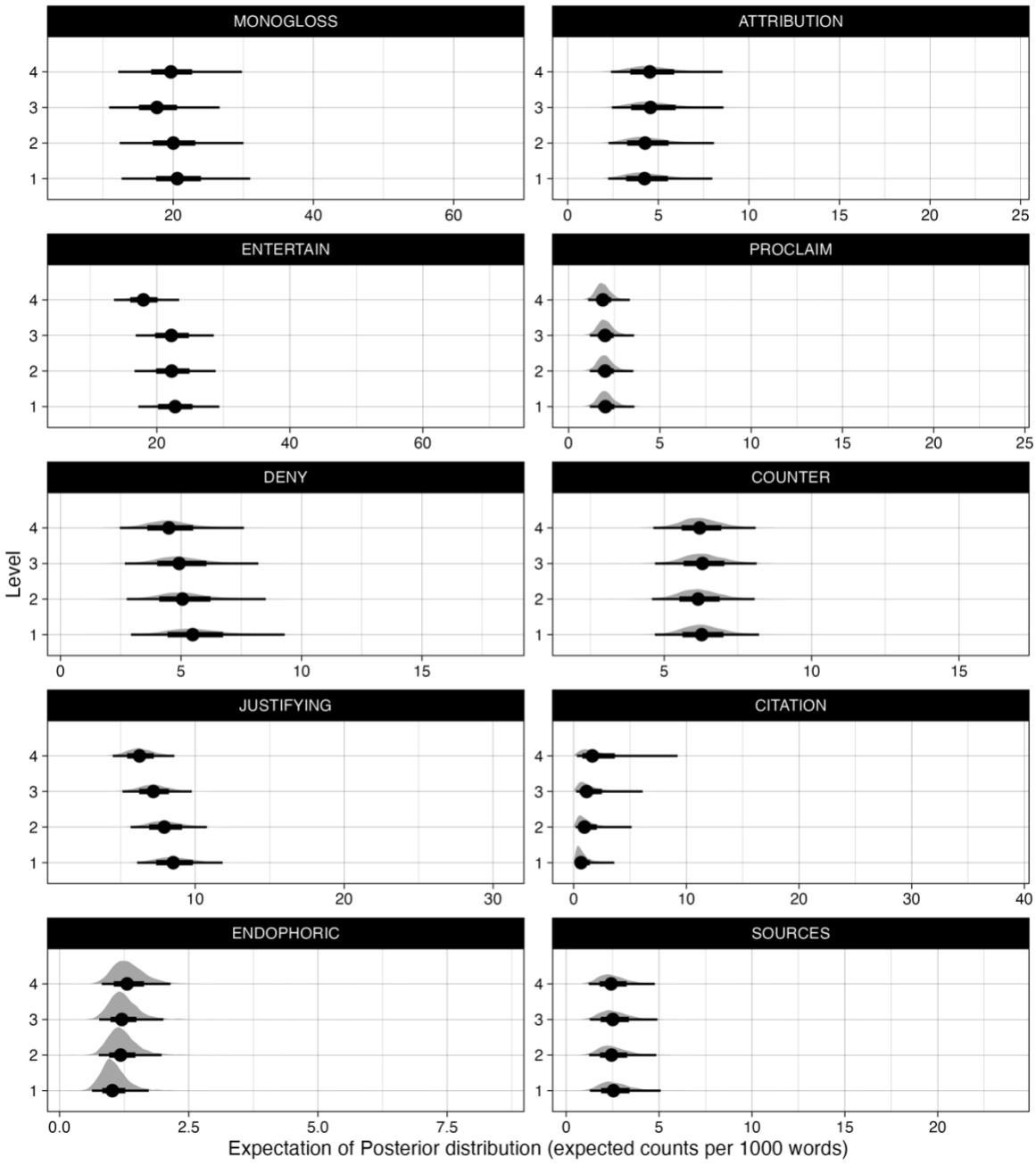
findings suggest that the frequency of each Engagement strategy alone may not be the most salient feature differentiated by assignment grades and course levels. Three crucial exceptions to this are the frequencies of CITATION, ENDOPHORIC, and ENTERTAIN. Figure 4.9 depicts a post hoc comparison of the frequencies of these categories across course levels.

The negligible effects of grades and course levels in most Engagement strategies differ from previous findings. Previous studies have shown that low-graded essays use MONOGLOSS more frequently, while high-graded essays tend to use a wider variety of Heteroglossic strategies (Lancaster, 2014; Wu, 2007). On the other hand, the results of this study indicate that the effects of grade or course level are minimal. These contradictory findings may be partly due to methodological differences. A key methodological difference is that the BAWE corpus mainly contains higher-performing writing (graded as “Merit” or “Distinction”, which are roughly equivalent to grades B and A, respectively).

In contrast, Lancaster (2007) and Wu (2017) selected their texts from a broader range of grades (C and A letter grades). Thus, the writing samples they analyzed may have highlighted differences more clearly. Relatedly, this study examined a corpus of academic English writing (BAWE) that were nearly 100 times larger than previous studies. This significantly larger sample may have produced more conservative, and arguably more accurate, estimates of the range of possible distributions of Engagement, which vary with assignment grades. Overall, regarding level differences, the current analysis of the BAWE corpus fails to replicate differences in the distributions of Engagement strategies according to levels. Future research should clarify whether the current non-significant patterns can be replicated by adding more lower-graded writing samples.

Figure 4.9

Expectations of the posterior distribution for marginal means by level.



Note. Predicted marginal means based on the multilevel MANOVA model are plotted. Model-based predictions of marginal means were generated by considering the effects of levels, while other random effects were integrated out from the original formula.

4.5.5 Pedagogical Implications

The present findings have several implications for pedagogy and the assessment of writing assignments in higher education. First, the large two-way interaction effect between genre family and discipline indicates that the register of university writing is a diverse phenomenon. Specifically, the findings indicate that rhetorical features of engagement in the critique genre family, for example, vary greatly depending on the discipline the student writes in. This finding has two important implications. First, content instructors may need to know that their expectations for the rhetorical features of writing for a particular genre family may differ from other instructors in adjacent fields. Similarly, instructors may need to know that a group of students are likely to have been trained in different approaches and do not share the same approach to writing the genre of the assignment. For this reason, both instructors' and students' awareness may be raised by presenting frameworks of the prototypical social and communicative purposes of each assignment, and could also be encouraged to use more specific category naming that reflects these communicative purposes rather than calling them *essays* for convenience. To this end, the BAWE corpus (Alsop & Nesi, 2009; Nesi & Gardner, 2012) and/ or MICUSP (Römer & O'Donnell, 2011; Römer & Swales, 2010) will be great resources for instructors and students alike in order to agree on the mutual expectation about what constitute a good writing in their own disciplines. In content courses in disciplines, specific communicative expectations may need to be clearly communicated to students so that instructors and students can build a mutual understanding of the requirements of given writing assignments.

Second, the same finding for discipline by genre family interaction has important implications for EAP programs and instructors as it rediscovers the benefits of raising students' awareness of the writing conventions of their disciplines, thus facilitating self-directed genre-

based learning (J. M. Swales & Feak, 2000, 2012). In this line of pedagogy, an EAP course can simultaneously cover both foundational rhetorical discourse features (e.g., cause and effect, compare and comparison, and definitions) while, at the same time, drawing students' attention to their own disciplines' unique rhetorical features. Such a pedagogical approach can be implemented in class by adapting the following generic activity instructions suggested by Swales and Feak (2012):

> Find and download two or three journal articles from your field that you think are well written. The articles do not necessarily have to be written by native speakers of English; however, they should be typical research articles in your field—not book reviews, editorial commentaries, or trade magazine articles from a publication with extensive advertising. […] Bring your articles to class so that you can reference them and gain an understanding of the writing conventions in your field. (p. 15)

Using this generic activity prompt, instructors can start to support students' own genre-based learning process. The instructions can be modified to target specific rhetorical, discourse features and/or the linguistic features required to realize them. For example, instructors can instruct students to focus on how engagement strategies (e.g., ATTRIBUTE, COUNTER, PROCLAIM) are used in their disciplinary writing. Such a focused awareness-raising activity may enhance student's understanding of how effective argumentation is structured from an engagement system perspective (see also Ryshina-Pankova, 2014). With such structured guidance from instructors, students may gain more confidence in exploring and understanding rhetorical features that are widely accepted or expected in their own field of study.

4.5.6 Limitations and future directions

This study has several limitations that should be addressed in future studies. First, although the current study used four different versions of the Engagement Analyzer and estimated uncertainties due to models through mixed-effect modeling, it is important to stress that the Engagement Analyzer pipeline needs some refinements. Notably, the current study indicates the possibility of CITATION tags being primarily affected by the format of writing (although I included different formats in the annotation dataset). Thus, more focused annotation of CITATION tags will help to refine the accuracy of analysis in the future. Nevertheless, the current study employed four separately trained NLP pipeline which showed relative strengths and weaknesses and treated them as a random effect in statistical analysis in order to avoid overconfidence in the results. It was also demonstrated that at the document level measures of frequency, the internal consistency of four models was almost perfect (inter-item correlations > .9; see Table 4.4). These results were also confirmed in the Multilevel MANOVA analysis (See Figure 4.4 for the non-significant effects of models). Thus, the selection of the particular model in the analysis is expected to be negligible in the current study.

In the current study, no comparisons were made between university written assignments and writing from different contexts (e.g., peer-reviewed journal articles, magazines and news articles). It is worth investigating, for example, how these university genres compare writing beyond university assignments, such as professional writing and business writing. This type of analysis will highlight how big cross-discipline or genre differences can be relative to writing in another context.

Third, this study has only considered the number of each Engagement strategy. Previous research suggests that not only the frequency but also specific sequences of Engagement

strategies may characterize the quality of an essay (Lancaster, 2014; Wu, 2007). As such, the qualitative insights gained in previous studies should be tested with quantitative measures of Engagement moving beyond sheer counts.

4.6 Conclusion

This study has sought to investigate whether and how university writing across different disciplines and genre families differs in terms of occurrences of engagement strategies. It also aimed to examine other variables related to assignments and writers, such as grades, levels, the writer's secondary education, and the writer's first language. The findings of a multilevel MANOVA indicate that frequencies of Engagement strategies vary significantly due to combinations of disciplines and genre families, suggesting that Engagement in university writing is more complex than simple characterization by genres or disciplines. The findings also suggest that while there is considerable variation in the distribution of engagement due to individual writers, higher grouping variables, such as writers' secondary education backgrounds and their L1, did not significantly affect their use of Engagement strategies. This study has also found that, contrary to previous findings (Lancaster, 2014; Wu, 2007), assignment grades were not a strong factor in differentiating the occurrence rates of Engagement strategies. These tentative findings need to be replicated with a broader range of writing assignments (especially those with lower grades) to tease apart methodological effects possibly affecting the current findings.

CHAPTER 5

STUDY 3: Engagement Strategies and Second Language Writing

5.1 Chapter Overview

In the preceding chapters, I reported the development and evaluation of the Engagement Analyzer (Chapter 3) and the application of the end-to-end NLP system to register analyses of university written assignments (Chapter 4). Both chapters demonstrate the benefits of an automatic approach to annotating rhetorical and discourse features of engagement (Martin & White, 2005). In the current chapter, I present a study that uses the Engagement Analyzer in another important domain of potential application—the analysis of written responses to a standardized L2 English proficiency exam. This area of application is important in the development of the Engagement Analyzer because it can inform the refinement of automated writing evaluation (AWE) systems frequently used in standardized assessment settings. In response to the recent call for the inclusion of rhetorical features in AWE systems to enhance construct coverage (Burstein et al., 2016; Lu, 2021), the current study tests the relative benefits of adding features of engagement strategies, in addition to existing lexical, syntactic, and cohesion features, in the prediction of writing scores on a large-scale standardized test of L2 English.

5.1.1 Measures of Heterogeneity and Evenness of Engagement strategies

Previous research suggests that lower-scoring essays are characterized by the dominance of MONOGLOSS strategies, with very few other heteroglossic strategies (See Chapter 2 for synthesis; also see Lancaster, 2014; Wu, 2007; Xu & Nesi, 2019). On the other hand, higher-scoring university assignments tend to exhibit a greater variety of Engagement strategies within a

document. To date, however, no quantitative measures of rhetorical diversity exist. To quantify the overall distribution of Engagement strategies, I have borrowed the concepts and quantitative measures of species heterogeneity and evenness from the field of ecology (Daly et al., 2018; Krebs, 1999a, 1999b; Morris et al., 2014). Heterogeneity and Evenness are two concepts (along with richness) that attempt to capture biodiversity in a specific community. Given a clear taxonomic classification (Krebs, 1999a), these concepts provide snapshots of distinct aspects of biodiversity in a given community. Richness is often conceptualized as the absolute number of species in a region (Krebs, 1999b, p. 534; Morris et al., 2014). This concept is analogous to the notion of type frequency in linguistics and the notion of abundance in the lexical diversity literature (e.g., Kyle et al., 2020). As the Engagement system provides a finite set of categories (at least according to the SFL literature, Martin & White, 2005), this concept may not be directly relevant in this study; however, the notions of heterogeneity and evenness (Krebs, 1999a) may well provide a means to capture the diversity of Engagement strategies. The following paragraph illustrates these two notions and how they are relevant to the calculation of diversity measures of Engagement strategies.

Figure 5.1 (adapted from Krebs, 1999/2013, p. 536) illustrates two different scenarios where the communities are different in terms of heterogeneity and/or evenness. In panel (b) of Figure 5.1, three different species of composition are plotted. Communities A and B are the same in richness as they contain exactly five species (i.e., types). Community C is more diverse in terms of richness. The notion of heterogeneity attempts to capture differences among the three communities; that is, between communities A and B, A should be higher in heterogeneity because when the number of species remains constant, the more evenly distributed the between species count is, the more heterogeneous the community will be (hence A is more heterogeneous than B).

As Krebs notes, the notion of heterogeneity captures two distinct processes—richness and evenness (Krebs, 1999a, p. 535). That is why, under the notion of heterogeneity, community C can be considered more diverse than B (because of the sheer richness of species in the community). Two oft-used measures of heterogeneity are Shannon's index and Simpson's index (Krebs, 1999a), meaning that these measures are known to confound the two processes—richness and evenness (Daly et al., 2018). For this reason, researchers attempt to separate evenness from measures of heterogeneity by correcting the measure for the maximal diversity that can be achieved with a given richness value. In this study, I employ one of the measures that is considered able to measure evenness—namely, Simpson's E index.

It is important to note, however, that strict distinctions between heterogeneity and evenness measures are not particularly relevant in the measurement of Engagement diversity because we know that the Engagement category is finite. Arguably, this naturally corrects Shannon's index and Simpsons's D index in order to focus on evenness among engagement strategies without any post hoc corrections. Thus, for this particular application, potential problems of classical diversity measures are minimized. For a more detailed discussion of the different processes that affect classical diversity indices, such as Shannon's and Simpson's indices and more recent alternatives, see Daly et al. (2018).

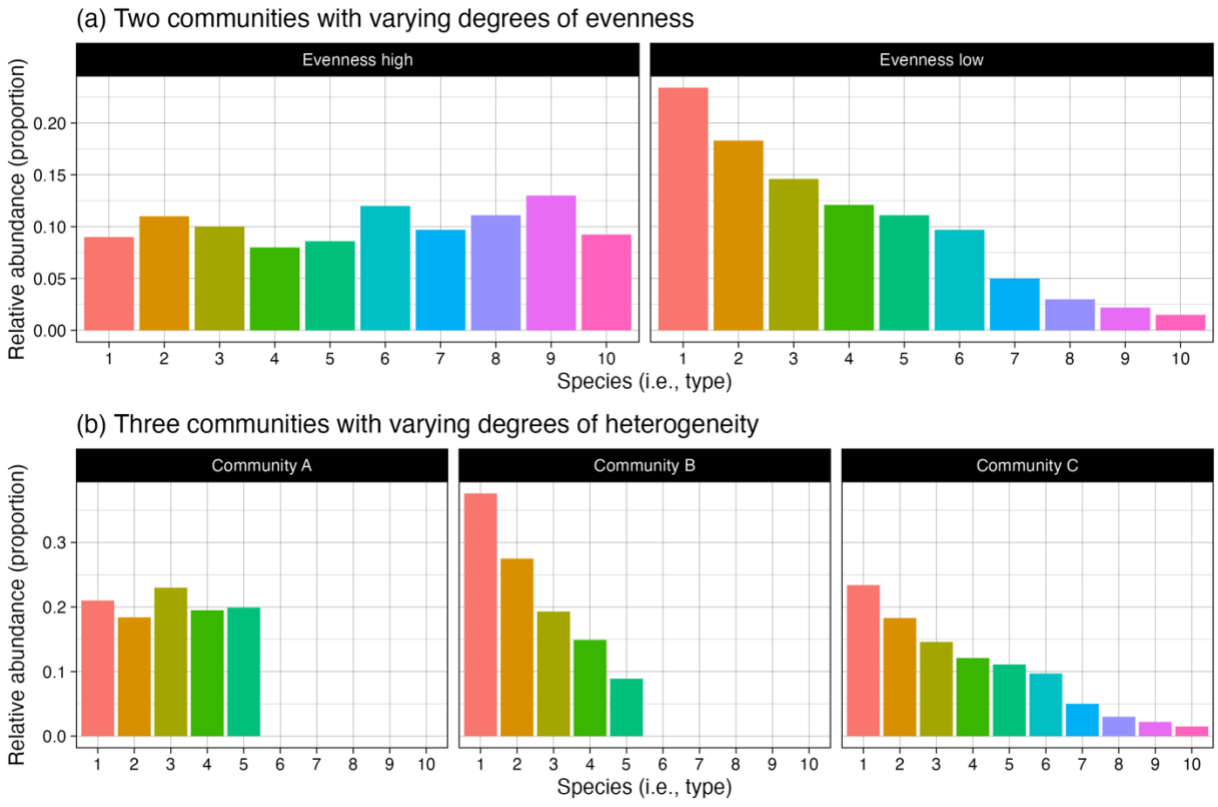
5.2 This Study

Given the paucity of research on how the use of engagement strategies influences timed L2 writing (cf. Lam & Crosthwaite, 2018; see Chapter 2 for the review), the current study investigates whether and how distributions of engagement strategies in writing performance are

related to writing scores across two task types. The study is guided by the following three research questions:

1. How are the rhetorical features of Engagement distributed across two writing task types: email and argumentative essay tasks?
2. What is the relationship between the rhetorical features of Engagement and essay qualities in argumentative and email writing tasks?
3. Do Engagement strategies explain second language essay qualities above and beyond existing linguistic features at the levels of lexis, syntax, and cohesion?

Figure 5.1
Illustrations of (a) Evenness and (b) Heterogeneity (adapted from Krebs, 1999/2013, p. 593).



5.3 Method

5.3.1 Examination for the Certificate of Competency in English (ECCE)

The data for this study came from a privately shared dataset for the Examination for the Certificate of Competency in English (ECCE) from the Michigan Language Assessment. ECCE is targeted as a standardized exam to certify that the test-taker has reached B2 level on the Common European Framework of Reference (CEFR; Council of Europe, 2020). ECCE assesses test-takers' English proficiency in four skill areas—Writing, Reading, Listening, and Speaking. A test-taker receives a scaled score from 0 to 1,000, with 650 or above considered having "passed" or reached CEFR B2 level.

In the writing section, test-takers have a choice of completing either an email writing task or an argumentative essay task. The email prompt consists of a situational description of the purpose of the email and the imagined audience. The essay prompt consists of a statement on which the test-taker is expected to take a stance. According to Michigan Language Assessment, the writing section is assessed by at least two certified raters on an analytic scale from 1 to 5 for four categories—(i) Content and Development, (ii) Organization and Connection of Ideas, (iii) Linguistic range, and Control, and (iv) Communicative effect. The shared dataset contains a scaled numerical score, summarized from raw scores on the four components.

5.3.2 Prompts and writing scores

As shown in Table 5.1, the shared dataset comprises a total of 698 samples distributed across two email prompts and two essay prompts. There are more essay responses (n = 398) than email responses (n = 298) in this dataset. Apart from the differences in sample size, one important difference between the two types of prompts relevant to this study is the contextual

parameters set by the two prompts. In light of the three-dimensional model of contextual parameters in Systemic Functional Linguistics (Halliday & Matthiessen, 2014), the two task types may be different in terms of tenor, which is related to the interlocutors, recipients, and audience of language use. In the context of this study, the email prompt requires test-takers to address an (imagined) recipient of the message, who typically has more perceived power relative to the writer (i.e., new school principal, city mayor) (cf. Brown & Levinson, 1987). On the other hand, in the argumentative essay writing task, the tenor is not explicitly specified.

Table 5.2 and Figure 5.1 present descriptive statistics for the writing scores across four prompts. Possibly because the scores are scaled, the writing scores do not significantly differ from each other $F_{trimmed\text{-}means}(3, 222.88) = 0.05$, $p = 0.99$. Out of 698 exam responses, 33 were illegible to the transcriber. The remaining 667 test scripts are used in the present study.

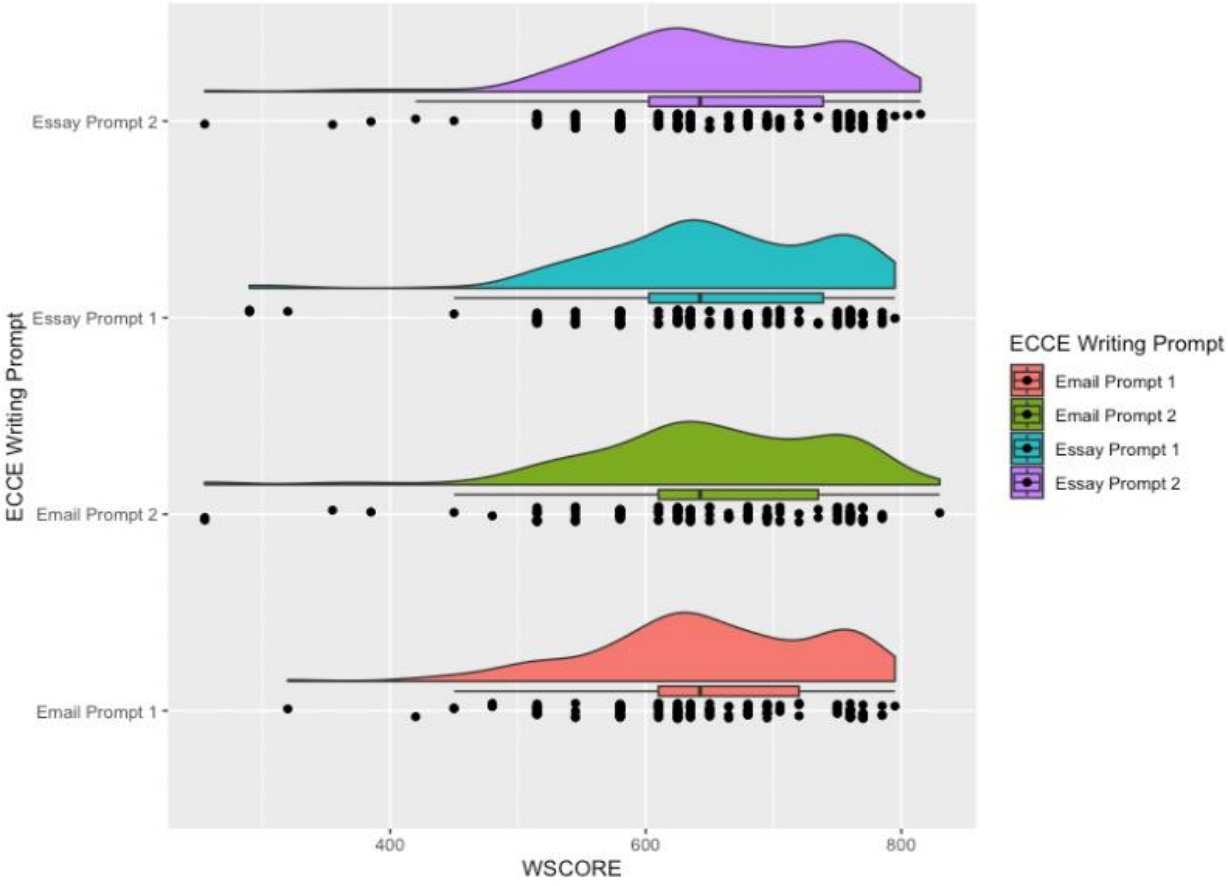
Table 5.1
Details of the four prompts.

Prompt ID	n	Communicative purpose/Topic
Email 1	150	Write an email to a new school principal, Ms. Wagner, to express your opinion about potential changes to school rules.
Email 2	148	Write an email to the city mayor, Mayor Sanchez, to express your opinion on what to build as a new public facility.
Essay 1	199	Airlines should make all seats smaller and stop offering in-flight services such as meals and entertainment in order to lower ticket prices.
Essay 2	199	All students in school should be required to learn how to play a musical instrument.

Table 5.2
Descriptive statistics for the writing scores across four prompts.

Prompt	*n*	*M*	*SD*	*Median*	*Min*	*Max*	*Skewness*	*Kurtosis*	*SE*
Email 1	148	648.986	89.008	635	320	795	-0.547	0.271	7.316
Email 2	143	651.294	98.216	650	255	830	-1.175	2.637	8.213
Essay 1	184	656.685	85.922	643	290	795	-0.747	1.497	6.334
Essay 2	192	655.885	91.099	650	255	815	-0.745	1.408	6.574
All	667	653.591	90.686	650	255	830	-0.832	1.645	3.511

Figure 5.2
Distribution of writing scores across four prompts.



5.3.3 Engagement Analysis

A total of 667 writing samples were analyzed via the Engagement Analyzer. As discussed in detail in Chapter 3, the Engagement Analyzer is an end-to-end machine-learning system trained on a human-annotated corpus, the Engagement Discourse Treebank (EDT; see Chapter 3). As demonstrated in Chapter 3, the accuracy of machine-learning systems tends to surpass the benchmarks between two human annotators, who major in Linguistics and were trained over the course of 10 weeks (over 50 hours of training time). However, it is also important to

acknowledge that the accuracy of a single model is still moderate-to-substantial (the average Cohen's Kappa = .72). For this reason, the current study employed four separately trained models from Chapter 3 to gain from the relative strengths of these models, which is analogous to hiring four different human coders for data analysis. Table 5.3 lists the performances of the four models on the test sets for each one (the same models as Study 2).

As can be seen from Table 5.3, three LSTM-based models (model [1]-[3]) and one Dual-Transformer model (model [4]) were considered. This choice was motivated by the relative strengths of these models as well as the distribution of held-out datasets they were trained on. Three versions of the best-performing LSTM model, same architecture and hyperparameter settings but trained on three versions of the held-out datasets, were included. Their by-tag performances were excellent, except for the ENDOPHORIC tag by model (2). To address this weakness, I included one version of the Dual-Transformer model, which performed exceptionally well on the ENDOPHORIC tag. It is important to stress that although the overall accuracy figures are similar to each other, each model has relative strengths. It is hoped that this demonstrates their strengths as well as the fact that they were trained and tested on different held-out datasets boosts the reliability of the overall analysis. The variabilities of the analyses by four different models were statistically controlled for through mixed-effect (or multilevel) modeling (see Section 5.3.6, Statistical analysis).

Table 5.3

Summary of models' performance used to identify engagement strategies (reproduced).

	LSTM model trained on Fold1			LSTM model trained on Fold3			LSTM model trained on Fold5			Dual Transformer model trained on Fold4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ATTRIBUTION	0.762	0.689	0.724	0.841	0.745	**0.791**	0.813	0.696	0.750	0.790	0.791	*0.790*
CITATION	0.965	0.976	**0.970**	0.912	0.932	0.922	0.948	0.964	*0.956*	0.904	0.899	0.901
COUNTER	0.836	0.963	**0.895**	0.903	0.836	0.868	0.836	0.888	0.861	0.921	0.854	*0.886*
DENY	0.950	0.884	**0.916**	0.830	0.835	0.833	0.863	0.838	0.850	0.854	0.886	*0.870*
ENDOPHORIC	0.703	0.723	0.713	0.528	0.563	0.545	0.900	0.619	0.734	0.858	0.837	**0.848**
ENTERTAIN	0.890	0.779	0.831	0.869	0.868	**0.869**	0.865	0.873	**0.869**	0.832	0.869	*0.850*
JUSTIFYING	0.862	0.821	*0.841*	0.957	0.758	**0.846**	0.819	0.790	0.804	0.785	0.825	0.805
MONOGLOSS	0.822	0.845	0.833	0.867	0.830	**0.848**	0.856	0.818	*0.837*	0.871	0.735	0.797
PROCLAIM	0.873	0.658	*0.750*	0.837	0.765	**0.799**	0.818	0.667	0.735	0.730	0.652	0.689
SOURCES	0.830	0.736	**0.780**	0.744	0.800	*0.771*	0.741	0.803	0.770	0.770	0.755	0.763
_	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
accuracy			0.741			*0.743*			**0.742**			0.741
macro avg	0.772	0.734	**0.750**	0.753	0.721	0.736	0.769	0.723	0.742	0.756	0.737	*0.745*
weighted avg	0.786	0.741	*0.761*	0.791	0.743	**0.765**	0.775	0.742	0.756	0.761	0.741	0.749
kappa			0.710			**0.712**			*0.711*			0.708

Note. F1 columns are highlighted for readability; the best F1 scores (by category) are bolded; second-best F1 scores (by category) are italicized; scores over .8 are highlighted in green; scores under .7 are highlighted in red.

5.3.4 Engagement measures

In this study, I use an essay as a unit of analysis. Two types of measures were computed for each piece of writing.

5.3.4.1 Normalized frequency counts

As with normalized frequency counts, the occurrence of each Engagement strategy is tallied, and they are normalized in the following formula:

$$Normed\ frequency\ per\ 100\ words = \left(\frac{Raw\ frequency}{Number\ of\ total\ words\ in\ piece\ of\ writing}\right) \times 100$$

This text-length normalization helps to reduce the confounding effect of essay length when predicting writing scores. This normalization formula is applied to all engagement strategies.

5.3.4.2 Evenness measures of Engagement strategies

As explained in the chapter overview, I calculate three measures that attempt to quantify the diversity of Engagement strategies in an essay. To calculate these measures, a vector of frequency counts for 10 engagement strategies is created for each essay.

5.3.4.2.1 Shannon's index H'.

Shannon's index of information entropy is a measure of heterogeneity (see Krebs, 1999b), and it is widely used in research on biodiversity (Morris et al., 2014). This measure is expressed in the following mathematical formula:

$$Shannon\ H' = -\sum_{i=1}^{S} p_i\ ln(p_i)$$

where S denotes the number of species in the sample, p_i is the proportion of i species in the sample. As a measure of entropy, Shannon's index measures the uncertainty of an individual belonging to a given category. Thus, this measure returns higher scores when the relative proportions of each category are even and/or there are more classes to classify (i.e., larger type counts in linguistic terms). For this reason, Shannon's index is technically a measure of heterogeneity (Krebs, 1999a), because it is also affected by the absolute number of species in the sample (Morris et al., 2014). However, I regard this as a measure of evenness in the current study because the Engagement system in this study provides a finite set. According to the literature, Shannon's index is classified as a Type I index, which tends to give roughly equal prominence to rare and abundant species (Daly et al., 2018) but is more sensitive to change in rare categories (Krebs, 1999a).

5.3.4.2.2 (Gini-)Simpson's index.

Simpson's index is a probabilistic measure of heterogeneity (see Daly et al., 2018), expressed in the following formula:

$$Simpson's\ D\ = \sum_{i\,=\,1}^{S} p_i^2$$

where S denotes the number of species in the sample, and p_i is the proportion of i species in the sample. Because a lower score on this original Simpson's measure is considered to reflect higher diversity, researchers often take its complement (1 – D) and call it Gini-Simpson's index (Jarvis, 2013b; Krebs, 1999b). In this study, I use Gini-Simpson's index to interpret a higher score to mean higher diversity. Gini-Simpson's index can be interpreted as "the probability that two individuals chosen at random will be different species" (Krebs, 1999b, p. 582). Krebs (1999/2013) classifies Gini-Simpson's index as Type II, which is more sensitive to a change in abundant species in the sample.

5.3.4.2.3 Simpson's Evenness.

One measure of evenness, Simpson's Evenness, is used in this study, which is calculated in the following formula:

$$Simpson's\ E\ = \frac{1\ /\ D}{S}$$

where D represents Simpson's D (see above). Note that the reciprocal of Simpson's original index (1 /D) is also frequently referred to as Simpson's dominance (Morris et al., 2014) or Hill's N_2 (Hill, 1973), which ranges from 0 to S. Thus, Simpson's evenness is considered to be a Simpson's dominance index normalized for the number of species in the sample. Accordingly, this index is more sensitive to changes in more dominant classes in the sample (Daly et al., 2018; Krebs, 1999a).

5.3.4.2.4 Reliability of Engagement measures

In order to examine the reliability of measures across different models of Engagement Analyzer, a series of intraclass correlation coefficients were computed. Intra-class correlation is a measure of internal consistency. It is calculated as the proportion of variance explained within a grouping variable (i.e., individual essay) over the total variance in the score. High intraclass correlation can be taken as evidence for the internal consistency of a measure. Table 5.4 shows that ICCs are satisfactory, ranging from .840 (ATTRIBUTION) to .970 (ENTERTAIN). The four measures of evenness also show internal consistency.

Table 5.4
Intraclass Correlations of Engagement measures.

Measures	ICC	LowerCI	UpperCI
ATTRIBUTION	0.840	0.821	0.858
COUNTER	0.964	0.959	0.968
DENY	0.948	0.941	0.954
ENTERTAIN	0.970	0.966	0.973
JUSTIFYING	0.942	0.935	0.949
MONOGLOSS	0.909	0.897	0.919
PROCLAIM	0.845	0.826	0.863
SOURCES	0.903	0.891	0.915
Shannon's H’	0.933	0.924	0.941
Simpson's D	0.920	0.910	0.930
Simpson's E	0.893	0.879	0.905

Note. ICC = Intraclass Correlation; CI = 95% Confidence Interval

5.3.5 Other linguistic measures

To answer Research Question 3, a series of existing linguistic measures were also considered. Their selection was motivated by previous studies showing the contribution of linguistic features to essay quality at the levels of lexis (Kim et al., 2018; Kyle et al., 2018), phraseology (e.g., Bestgen & Granger, 2014; Kyle & Eguchi, 2021), syntax (Kyle & Crossley, 2017, 2018; Lu, 2010; Wolfe-Quintero et al., 1998), and cohesion (e.g., Crossley et al., 2016, 2019). These domains of linguistic features as well as the individual indices considered are reviewed below.

5.3.5.1 Lexical features—diversity, sophistication, and phraseological sophistication

Table 5.5 lists a total of 14 indices used to construct lexical baseline models.

As for lexical diversity, Jarvis (2013) lists six theoretical subconstructs that comprise the notion of lexical diversity—size, richness, effective number of types, evenness, disparity, importance, and dispersion. However, the most frequently used sub-construct of lexical diversity

is still lexical variety, measured with variants of type-token ratio. Following recent studies on the reliability and validity of lexical variety measures (e.g., Kyle, Crossley, et al., 2021; Zenker & Kyle, 2021), Moving Average Type-Token Ratio (MATTR; Covington & McFall, 2010) was selected. MATTR is calculated using the Tool for the Automatic Analysis of Lexical Diversity (TAALED; Kyle, Crossley, et al., 2021).

Research has demonstrated the advantages of a multivariate approach to lexical sophistication (Eguchi & Kyle, 2020; Kim et al., 2018; Kyle & Eguchi, 2021). This line of research has found that high-performing essays are characterized by a variety of lexical characteristics, including but not limited to: (a) word frequency, (b) concreteness (Brysbaert et al., 2014), (c) age of acquisition (Kuperman et al., 2012), and (d) contextual distinctiveness (McDonald & Shillcock, 2001). All measures are computed using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018; Kyle & Crossley, 2015).

5.3.5.2 Fine-grained syntactic complexity and sophistication

A total of 15 measures of fine-grained syntactic complexity and sophistication are considered in this study (Table 5.6). The selection of measures was based on previous studies demonstrating the benefits of using fine-grained clausal and phrasal measures (Kyle & Crossley, 2018) and verb-argument construction indices (Kyle & Crossley, 2017). In particular, in selecting clausal and phrasal features, I tried to balance the degree of graduality of measures for proficiency levels represented in the ECCE corpus based on developmental progression reported in the previous study. As such, finer-grained clausal features are included (e.g., clausal complement, adjectival complement, to-infinitive clauses, etc.) while coarser-grained nominal dependency measures (e.g., average number of nominal dependents and their standard deviations) are used instead of finer-grained ones. All measured are computed using the Tool for

the Automatic Analysis of Syntactic Complexity and Sophistication version 1.3.8 (TAASSC; Kyle & Crossley, 2017, 2018).

Table 5.5.
Lexical, phraseological measures in the lexical baseline model.

Domain	Predictors
Lexical diversity	Moving Average TTR (50 words) All words
Lexical Sophistication	Kuperman Age of Acquisition, Content words
Lexical Sophistication	Kuperman Age of Acquisition, Function words
Lexical Sophistication	Brysbaert Concreteness, Content Words
Lexical Sophistication	Brysbaert Concreteness, Function Words
Lexical Sophistication	COCA Magazine Frequency, Log Content words
Lexical Sophistication	COCA Magazine Range, Content words
Lexical Sophistication	McDonald Contextual Diversity, Content words
Lexical Sophistication	ALL Academic Word List
Phraseological Sophistication	ALL Academic Formulas List
Phraseological Sophistication	COCA Magazine Bigrams, Mutual Information
Phraseological Sophistication	COCA Magazine Bigrams, Delta P
Phraseological Sophistication	COCA Magazine Trigrams, Mutual Information
Phraseological Sophistication	COCA Magazine Trigrams, Mutual Information2

Table 5.6
Fine-grained syntactic dependency and verb-argument construction measures in the syntax baseline model.

Domain	Predictors
Fine-grained syntactic Complexity	dependents per nominal
Fine-grained syntactic Complexity	dependents per nominal (standard deviation)
Fine-grained syntactic Complexity	adjective complements per clause
Fine-grained syntactic Complexity	adverbial clauses per clause
Fine-grained syntactic Complexity	clausal complements per clause
Fine-grained syntactic Complexity	clausal subjects per clause
Fine-grained syntactic Complexity	nominal subjects per clause
Fine-grained syntactic Complexity	passive nominal subjects per clause
Fine-grained syntactic Complexity	prepositions per clause
Fine-grained syntactic Complexity	open clausal complements per clause
Fine-grained syntactic Complexity	adverbial modifiers per clause
Fine-grained syntactic Complexity	modal auxiliaries per clause
Verb-Argument Construction	average lemma construction combination frequency, log -transformed – all
Verb-Argument Construction	average delta p score verb (cue) – construction (outcome) (types only) – all
Verb-Argument Construction	average delta p score construction (cue) – verb (outcome) (types only) – all

5.3.5.3 Cohesion features

A total of 13 indices were used to create the cohesion baseline model (Table 5.7). There are four indices related to lexical cohesion. The two synonym overlap measures (noun and verb) were calculated based on the synset in the WordNet database (Miller, 1995), calculating the number of words from the same synset between paragraphs. The other two lexical cohesion measures use more recent vector semantic models to measure the similarities in lexical use between two or more adjacent paragraphs. According to Crossley et al. (2019), word2vec similarity all (Paragraph) 1 calculates the semantic similarity between two adjacent paragraphs; word2vec similarity all (Paragraph) 2 calculates the semantic similarity between one paragraph and the following two paragraphs. I also included a total of seven connective indices partly because they are conceptually related to a subset of engagement items (e.g., *however*, *although*).

Table 5.7
Lexical cohesion and connective measures in the cohesion baseline model.

Domain	Predictors
Lexical cohesion	Synonym overlap Paragraph (noun)
Lexical cohesion	Synonym overlap Paragraph (verb)
Lexical cohesion	word2vec similarity all (Paragraph) 1
Lexical cohesion	word2vec similarity all (Paragraph) 2
Connectives	Basic connectives
Connectives	Conjunctions
Connectives	Addition
Connectives	Sentence Linking
Connectives	Order
Connectives	Reason and Purpose
Connectives	All Causal
Connectives	Positive Causal
Connectives	Opposition

5.3.6 Statistical analyses

All the main statistical analyses were conducted within a Bayesian framework (Gelman et al., 2015; McElreath, 2020) using the *brms* package (Bürkner, 2017), which uses stan for backend computation (Gelman et al., 2015). As already explained in detail in Chapter 4, an advantage of Bayesian analysis is that it includes a more intuitive probabilistic expression of estimated parameter values and their uncertainties (Dienes, 2011; McElreath, 2020; Winter & Bürkner, 2021).

5.3.6.1 RQs 1 and 2: Distributions of rhetorical strategies of Engagement across task types and their relationship with assessed essay quality

To answer RQs 1 and 2, I fit a series of multivariate Poisson regressions. As in Study 2 (see Chapter 4), Poisson regression is deemed appropriate considering the nature of the outcome variable (i.e., raw frequency counts of Engagement strategies). To investigate the effects of task-type—argumentative essay versus email writing—on the distribution of Engagement strategy (RQ1) as well as the relationship between writing score and Engagement (RQ2), a single-level Poisson regression was fitted with three fixed effects predictors—(a) Task type, (b) Writing score, and (c) two-way interaction between (a) and (b).

```
brm(mvbind( MONOGLOSS, ATTRIBUTION, ENTERTAIN, JUSTIFYING, COUNTER, DENY, PROCLAIM) ~
        scale(WSCORE) * Genre
        + (1|model)
        + offset(log(nwords_200)),
        family = poisson(link = "log"),
          data = ECCE_data_long,
          cores = 4,
          chains = 4,
        iter = 4000,
           prior = c(set_prior("student_t(3,0,1)", class = "b", resp = resp),
                  set_prior("student_t(3,0,1)", class = "sd", resp = resp)),
        backend = 'cmdstanr',
        file_refit = 'on_change',
```

```
file = 'models/ml_mnom0.1',

control= list(adapt_delta = .90))
```

In this formulation, essay length essay normalized for 200 words is used as the offset term to control for any effects of text length in the estimation of the occurrence of Engagement strategies. The binary task type is centered so that the interpretation of the intercept reflects the grand mean. Writing score is also centered using a z-score so that the intercept as well as the main effect of task type can be interpreted as conditioned on the mean of writing score. Two-way interaction estimates differences in the slopes of writing scores on the distributions of Engagement strategies due to task types. A set of weakly informative prior distributions was used. Specifically, for the fixed effects parameters and random intercept, I used student-t distributions with 3 degrees of freedom, a mean of 0, and a scale of 1, following the general guidelines from the Stan Development Team (2022). A student-t distribution was used instead of a normal distribution because the level of the factor was very low, and a conservative estimate was deemed appropriate in the current design.

To supplement the interpretation of the above model, the multivariate Poisson regression was expanded. This second model included writing scores as the only fixed effect. Instead, I also included varying intercepts of each category for Prompt, and varying slopes of writing scores in each category for each Prompt.

```
brm(mvbind( MONOGLOSS, ATTRIBUTION, ENTERTAIN, JUSTIFYING, COUNTER, DENY, PROCLAIM) ~
                scale(WSCORE) + offset(log(nwords_200)) +
                      (1 + scale(WSCORE)|| Prompt) +
                      (1 | model),
                    family = poisson(link = "log"),
                    data = ECCE_data_long,
                    cores = 4,
                    chains = 4,
                    iter = 5000,
                    backend = 'cmdstanr',
                    file_refit = 'on_change',
                    file = 'models/ml_poisson_RQ2.2',
                    control = list(adapt_delta = .93))
```

In this statistical formulation, the population effect (i.e., so-called fixed effect) of writing score expresses the grand mean effects of writing score on the distributions of Engagement strategies. Any varying slopes of writing score estimate by-prompt variabilities in the relationships between writing score and the distribution of Engagement strategies. This formulation allows statistical tests on each of the by-prompt slopes of writing scores.

5.3.6.2 RQ3: Combined model to predict essay quality

A series of multilevel linear regression analyses were conducted to investigate whether rhetorical features of Engagement contribute to essay quality above and beyond existing linguistic measures at the levels of lexis, phraseology, syntax, and cohesion. Specifically, I constructed a total of 16 regression models with different subsets of predictors, and I compared the performance of these competing models using information criteria that assess relative fit to the data. This model comparison approach is often preferred to a variable selection approach (McElreath, 2020). This is because while variable selection (e.g., stepwise regression) is conducted purely from a statistical perspective and ignores the theoretical implications of retained measures, a model comparison allows comparisons of alternative hypotheses expressed in separate regression models more directly. Since I am interested in whether the addition of rhetorical features of Engagement improves the model's predictions above and beyond existing lexical, phraseological, syntactic, and cohesion measures, I planned four sets of model comparisons. These blocks of model comparisons were between Engagement measures versus (a) lexical and phraseology features, (b) syntactic features, (c) cohesion features, and (d) all three domains of linguistic features. For each of the model comparison blocks, I constructed a total of four regression models to compare one to another:

1. Baseline linguistic features model: Regression model with linguistic features only
2. Linguistic features + Engagement diversity
3. Linguistic features + Engagement diversity and normed frequencies of each Engagement category
4. Engagement diversity and normed frequency measures

In each of the model comparison blocks, if measures of the rhetorical features of Engagement contribute to the prediction of writing scores, Models 2, 3, and/or 4 will outperform Model 1. Otherwise, the contribution of Engagement features can be considered minimal in predicting writing scores. I employed weakly informative priors for each of the analyses (Gelman, 2016). Specifically, I used a student-t distribution with 3 degrees of freedom, a mean of 0, and a scale of 1 on standardized regression parameters. For any group-level effects (i.e., random effects) I used a half-student-t distribution with 3 degrees of freedom, a mean of 0, and a scale of 2.5. For the intercepts, I used default priors estimated by the *brms* package to avoid specifying extreme prior values that might inadvertently affect posterior estimation. For model comparisons, I used the compare_performance() function in the performance package (Lüdecke et al., 2021).

Since Bayesian posterior distribution is difficult to estimate analytically, Markov chain Monte Carlo (MCMC) sampling was conducted using the *brms* (Bürkner, 2017), and *cmdstanr* (Gabry & Češnovar, 2021) packages. In Bayesian posterior inferences, it is important to check whether each chain of MCMC converges to the same regions of parameter values. Convergence is checked via the inspection of R-hat values and a visual inspection of trace plots (Vehtari et al., 2020). Specifically, R-hat ≤ 1.01 is interpreted as posing no problems in convergence. It is also important to ascertain how much impact the selection of prior distributions has on posterior estimation. To this end, I conducted a series of sensitivity analyses by varying the selection of prior distributions.

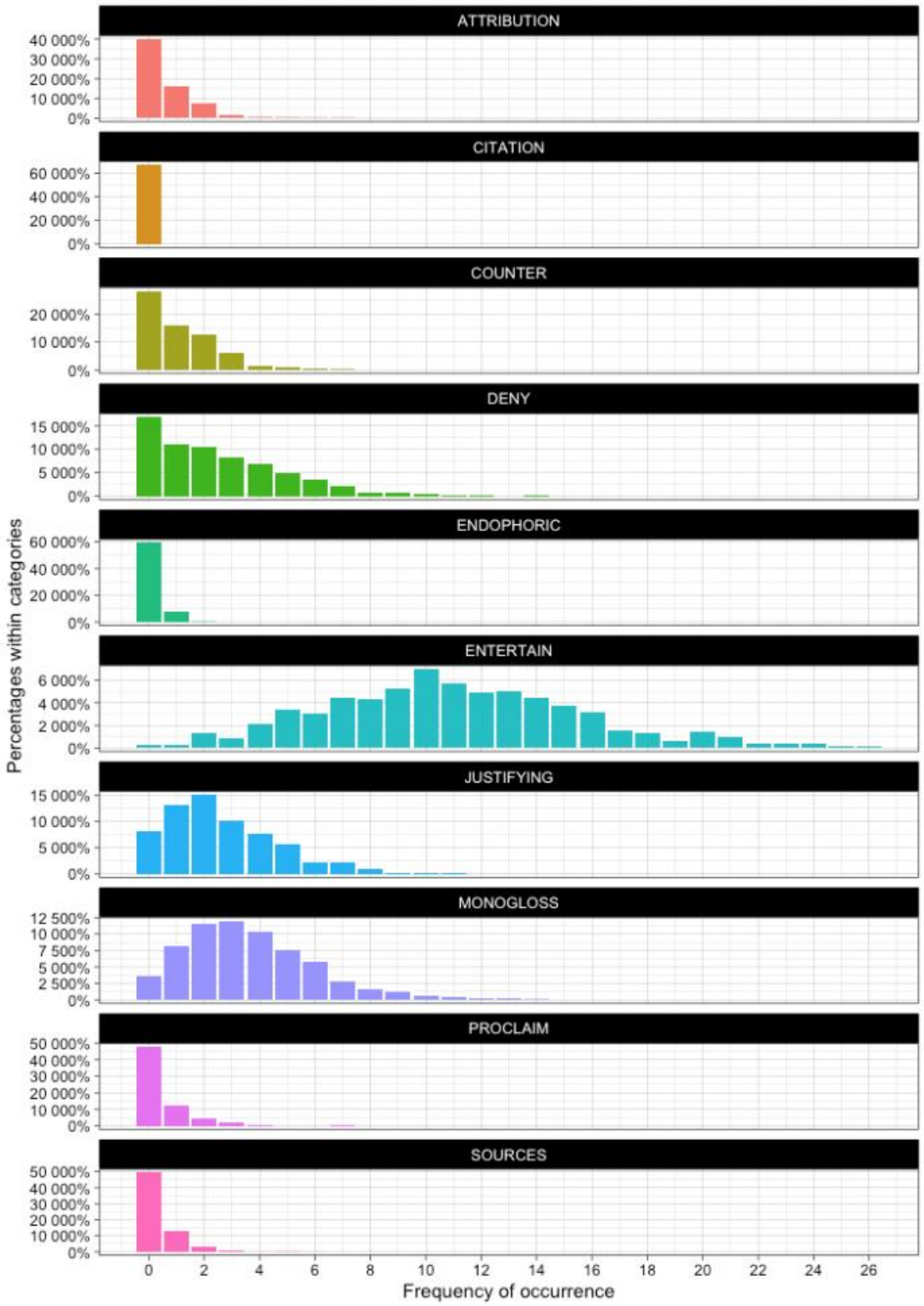
5.4 Results

5.4.1 Preliminary Analyses

Figure 5.3 summarizes the raw frequency counts of Engagement strategies with the number of times each strategy occurs in a document on the x-axis, and the relative proportions of the particular numbers of occurrences on the y-axis. A few observations can be made. First, the most frequent category overall was ENTERTAIN strategy. The mode of this distribution was 10 times per writing. This was followed by MONOGLOSS and JUSTIFYING, whose most likely occurrences were 3 or 4 times per document. As one might expect, there are no occurrences of CITATION and very few ENDOPHORIC references, as this is a timed essay approximately 200 words in length. ENTERTAIN strategy showed the most spread, with as few as 0 occurrences per 26 times at most. Since CITATION, ENDOPHORIC, and SOURCES tags were extremely rare in the timed L2 writing, they were not considered in the subsequent analysis.

Figure 5.3

Raw frequency counts of Engagement strategies per exam response.



5.4.2 RQ 1: Distributions of rhetorical strategies of Engagement across task types

To examine whether the two task types—email and argumentative writing—differ in their distribution of Engagement strategies, a multi-level Poisson regression was fitted. Table 5.8 presents the results of this analysis. The results indicate that meaningful task type differences were observed for all categories but ENTERTAIN. For example, the use of MONOGLOSS was 1.32 times more frequent in Emails than in Essays (i.e., exp[0.28]). In contrast, four heteroglossic categories were more frequent in Essay writing tasks—that is, ATTRIBUTION, JUSTIFYING, COUNTER, DENY, and PROCLAIM being 2.05, 1.23, 1.80, 2.03, and 1.16 times more frequent in Essay tasks than in Email, respectively. This result suggests that the two task types likely differ in the distributions of engagement strategies.

This pattern was confirmed through additional regression analysis based on prompt as the predictor rather than task type. The model with prompt as predictor allows further examination of both within and between task type differences in the frequencies of engagement strategies. Figure 5.4 displays a series of interval estimates of marginal means across prompts based on the second regression model. A few important convergent and divergent patterns are worth noting. First, as indicated in the first regression model, between task type differences were not very pronounced for ENTERTAIN, and PROCLAIM as the Credible Intervals (CrIs) overlap both within and between task type comparisons. This indicates that the distinction between email and argumentative may not be a principal motivating factor for selecting these Engagement strategies. Second, the DENY category showed a unique pattern whereby one of the email prompts contained many fewer DENY strategies. This observed task-type difference may be amplified due to these prompts. This probably shows the proposition-sensitive nature of the DENY category.

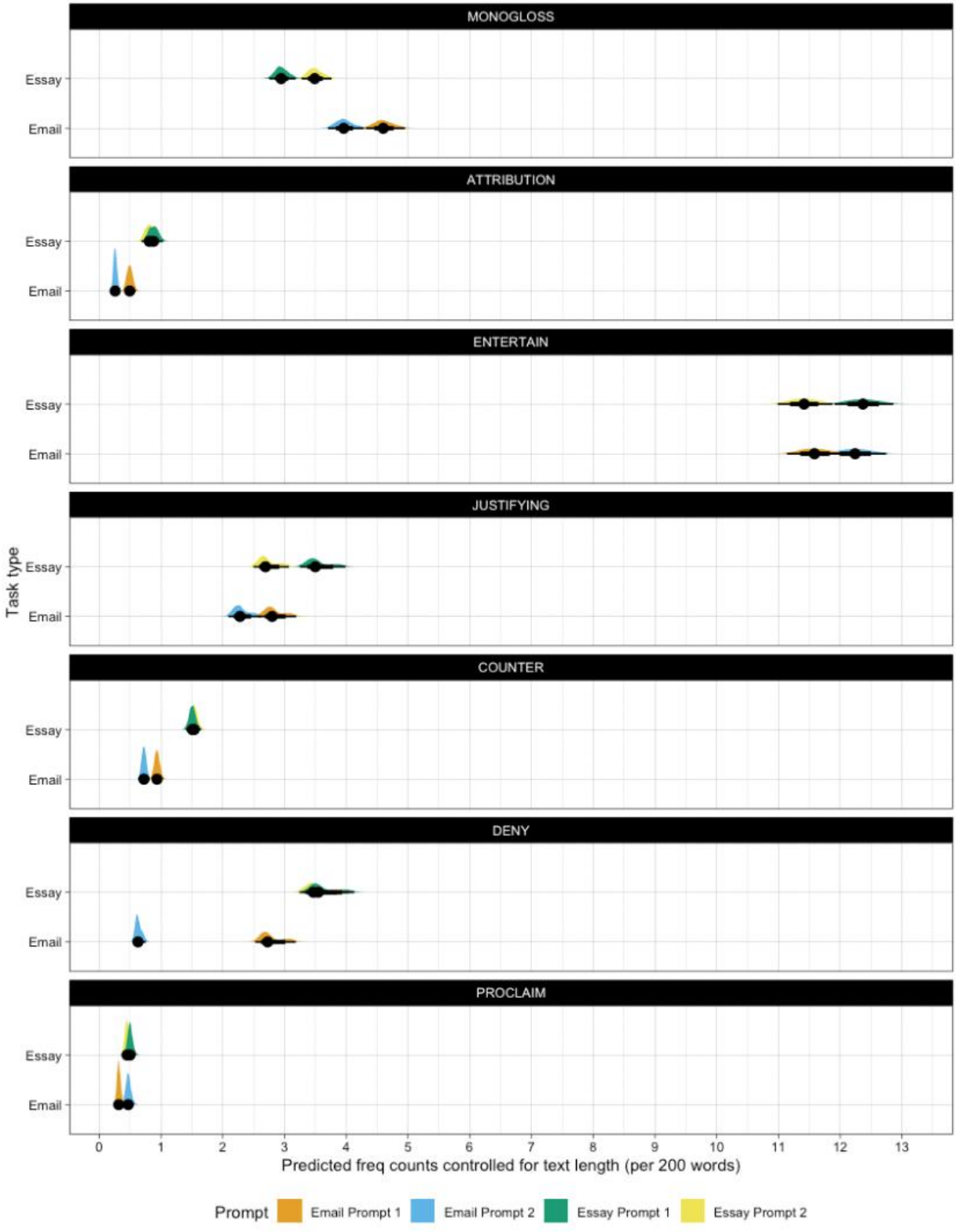
Table 5.8

Summary of single-level Poisson model predicting Engagement strategies.

Predictors	Estimate	Est.Error	l-95%	u-95%	Rhat	Bulk_ESS	Tail_ESS
MONOGLOSS							
Intercept	1.33	0.04	1.22	1.41	1.00	1517	1011
scaleWSCORE	-0.15	0.01	**-0.17**	**-0.13**	1.00	6219	2994
Genre1	0.28	0.02	**0.23**	**0.31**	1.00	7138	2862
scaleWSCORE:Genre1	-0.01	0.02	-0.05	0.03	1.00	8166	2860
ATTRIBUTION							
Intercept	-0.63	0.1	-0.86	-0.43	1.00	1560	1701
scaleWSCORE	0.19	0.03	**0.13**	**0.25**	1.00	6377	2862
Genre1	-0.72	0.06	**-0.83**	**-0.6**	1.00	5534	2998
scaleWSCORE:Genre1	-0.21	0.06	**-0.33**	**-0.09**	1.00	5219	2798
ENTERTAIN							
Intercept	2.47	0.03	2.42	2.52	1.00	1239	943
scaleWSCORE	0.03	0.01	**0.02**	**0.04**	1.00	7813	2846
Genre1	-0.01	0.01	-0.04	0.01	1.00	6351	2782
scaleWSCORE:Genre1	0.08	0.01	**0.06**	**0.11**	1.00	6264	3063
JUSTIFYING							
Intercept	1.05	0.07	0.91	1.19	1.00	1224	1041
scaleWSCORE	-0.06	0.01	**-0.08**	**-0.04**	1.00	6808	2732
Genre1	-0.21	0.02	**-0.26**	**-0.17**	1.00	6505	3061
scaleWSCORE:Genre1	0.14	0.03	**0.09**	**0.2**	1.00	6490	2613
COUNTER							
Intercept	0.06	0.04	-0.01	0.12	1.00	2057	1095
scaleWSCORE	0.22	0.02	**0.18**	**0.26**	1.00	6072	3201
Genre1	-0.59	0.04	**-0.67**	**-0.51**	1.00	5112	2618
scaleWSCORE:Genre1	-0.01	0.04	-0.1	0.08	1.00	5154	3137
DENY							
Intercept	0.92	0.08	0.77	1.1	1.00	1597	1332
scaleWSCORE	-0.06	0.01	**-0.08**	**-0.03**	1.00	6162	2948
Genre1	-0.71	0.03	**-0.76**	**-0.65**	1.00	6276	2798
scaleWSCORE:Genre1	-0.15	0.03	**-0.2**	**-0.1**	1.00	5472	2828
PROCLAIM							
Intercept	-0.95	0.09	-1.13	-0.77	1.00	1542	1261
scaleWSCORE	0.33	0.04	**0.26**	**0.4**	1.00	5479	2899
Genre1	-0.15	0.07	**-0.29**	**-0.01**	1.00	5870	3197
scaleWSCORE:Genre1	-0.06	0.07	-0.21	0.08	1.00	5312	3309

Figure 5.4

Predicted frequency counts of seven Engagement strategies across task types and prompts.



5.4.3 RQ 2: Relationship between rhetorical strategies and essay quality

To answer Research Question 2, the regression model presented earlier in Table 5.5 was re-examined for the remaining predictors—writing score and two-way interaction between writing score and task type. The 95% Credible Intervals of the two-way interaction terms did not exclude zero for MONOGLOSS (95% CrI = [-0.05; 0.03]), COUNTER (95% CrI = [-0.1; 0.08]), and PROCLAIM (95% CrI = [-0.21; 0.08]). This means that the relationships between writing scores and the frequencies of these categories were not different due to task type. Further, the main effects of writing scores for these three categories excluded zero in their CrIs, suggesting that writing scores were related to the use of these Engagement strategies, regardless of the choice of task type. Specifically, all slopes except MONOGLOSS were positive, suggesting that a higher writing score involves more frequent use of COUNTER, and PROCLAIM across the task type. The 95% Credible Interval for interaction terms excluded zero for ATTRIBUTION, DENY, JUSTIFYING, and ENTERTAIN. This means that there were differences in the relationships between writing scores and the occurrences of these four Engagement strategies across task types. Specifically, the positive interaction effects of ENTERTAIN and JUSTIFYING mean that the slopes for email prompts tended to be more positive than for Essay writing.

To further illustrate specific patterns of interaction, another multilevel ANOVA with a Poisson response distribution was fitted. Here, I reformulated the regression model to include varying intercepts of prompts for each Engagement strategy response and varying slopes of scores across prompts to model both grand mean effects and variabilities around that mean due to prompts. To examine by-prompt slopes, posterior samples were drawn for each slope (Table 5.9). Figures 5.5–11 present these slope estimates. Focusing on ENTERTAIN, JUSTIFYING, and DENY, whose fixed effects estimates with the first regression did not indicate a clear tendency,

the second regression with multilevel formulation gave more specific answers. For example, the 95% CrIs for each slope show that the use of ENTERTAIN were positively related to writing scores on email writing tasks, but this relation was negative or not meaningful in the Essay tasks. The patterns for DENY paint a complicated picture, where the use of DENY decreased as a function of writing scores in three tasks (Email 1 and 2 and Essay 1).

Overall, the results indicate that the use of Engagement strategies was associated with writing score, suggesting the potential of these rhetorical features to explain writing score above and beyond existing linguistic measures. This is the aim of RQ3.

Table 5.9

Summary of posterior draws of by-Prompt slopes estimating the effects of writing score on the occurrence of each Engagement category.

Response Category	Prompt	Estimate	Est.Error	CrI.Lower	CrI.Upper	1 – PD
MONOGLOSS	Email 1	-0.1023	0.0199	**-0.1348**	**-0.0689**	**< .001**
MONOGLOSS	Email 2	-0.1974	0.0196	**-0.2293**	**-0.1646**	**< .001**
MONOGLOSS	Essay 1	-0.1047	0.0230	**-0.1424**	**-0.0658**	**< .001**
MONOGLOSS	Essay 2	-0.1737	0.0193	**-0.2055**	**-0.1420**	**< .001**
ATTRIBUTION	Email 1	0.2577	0.0671	**0.1485**	**0.3707**	**< .001**
ATTRIBUTION	Email 2	-0.1334	0.0820	-0.2656	0.0026	0.0527
ATTRIBUTION	Essay 1	0.2583	0.0477	**0.1795**	**0.3372**	**< .001**
ATTRIBUTION	Essay 2	0.3213	0.0476	**0.2439**	**0.4013**	**< .001**
ENTERTAIN	Email 1	0.0358	0.0131	**0.0144**	**0.0573**	**0.0020**
ENTERTAIN	Email 2	0.1015	0.0132	**0.0801**	**0.1234**	**< .001**
ENTERTAIN	Essay 1	-0.0220	0.0119	**-0.0416**	**-0.0023**	**0.0323**
ENTERTAIN	Essay 2	-0.0021	0.0114	-0.0206	0.0170	0.4240
JUSTIFYING	Email 1	0.0063	0.0263	-0.0364	0.0495	0.4088
JUSTIFYING	Email 2	0.0208	0.0289	-0.0263	0.0685	0.7643
JUSTIFYING	Essay 1	-0.1942	0.0209	**-0.2290**	**-0.1610**	**< .001**
JUSTIFYING	Essay 2	-0.0571	0.0231	**-0.0953**	**-0.0198**	**0.0071**
COUNTER	Email 1	0.2227	0.0369	**0.1622**	**0.2843**	**< .001**
COUNTER	Email 2	0.2183	0.0406	**0.1494**	**0.2837**	**< .001**
COUNTER	Essay 1	0.2498	0.0345	**0.1978**	**0.3109**	**< .001**
COUNTER	Essay 2	0.2005	0.0303	**0.1481**	**0.2475**	**< .001**
DENY	Email 1	-0.1136	0.0251	**-0.1540**	**-0.0722**	**< .001**
DENY	Email 2	-0.1107	0.0499	**-0.1926**	**-0.0292**	**0.0141**
DENY	Essay 1	-0.0688	0.0212	**-0.1028**	**-0.0339**	**< .001**
DENY	Essay 2	0.0942	0.0217	**0.0587**	**0.1296**	**< .001**
PROCLAIM	Email 1	0.4881	0.0934	**0.3391**	**0.6474**	**< .001**
PROCLAIM	Email 2	0.1661	0.0751	**0.0443**	**0.2949**	**0.0104**
PROCLAIM	Essay 1	0.3743	0.0630	**0.2733**	**0.4782**	**< .001**
PROCLAIM	Essay 2	0.3502	0.0609	**0.2509**	**0.4499**	**< .001**

Note. CrI = 95% Credible Intervals; PD = Probability of direction, indicating the probability of a posterior distribution in one direction; the complement of PD (1 – PD) is considered analogous to a frequentist p-value in the sense that it can show the probability of the posterior not crossing zero.

Figure 5.5
Slopes of writing score on MONOGLOSS strategy by prompt.

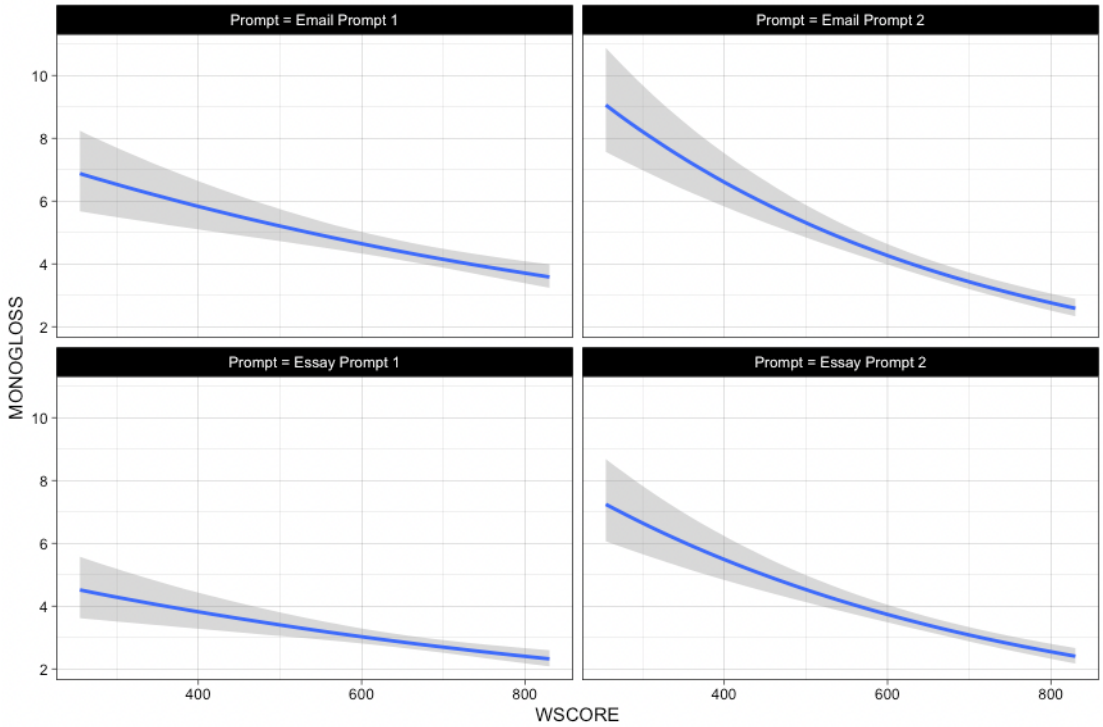


Figure 5.6
Slopes for writing score and ATTRIBUTION strategy by prompt.

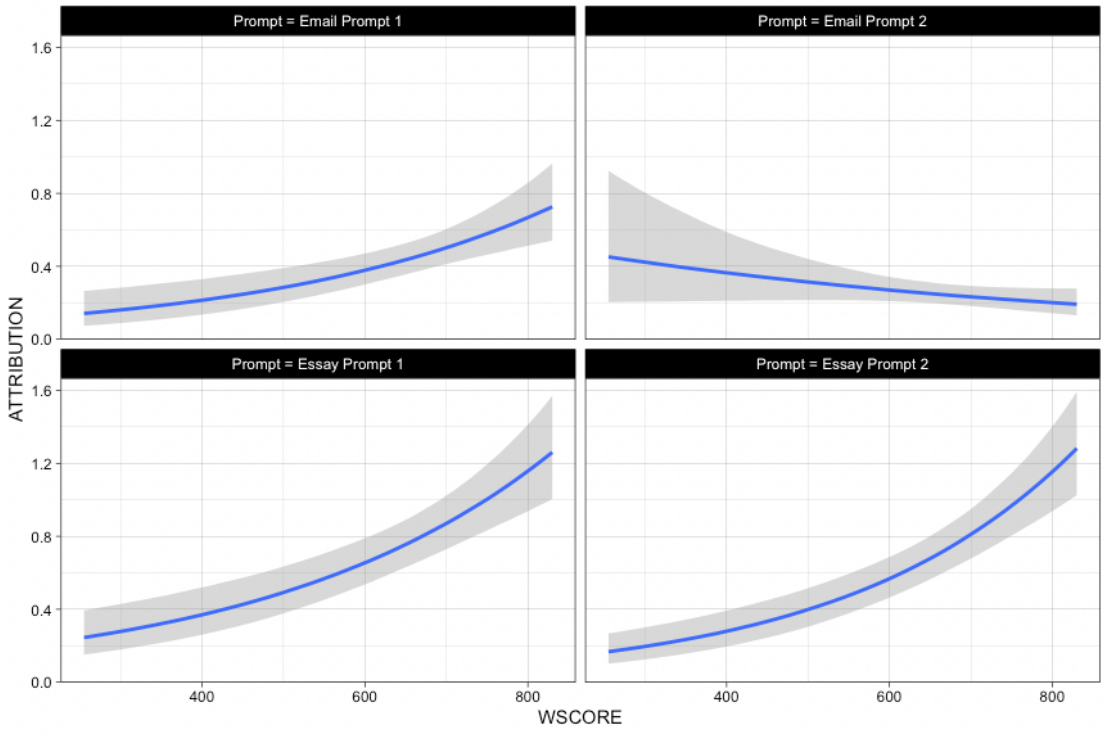


Figure 5.7

Slopes for writing score and ENTERTAIN strategy by prompt.

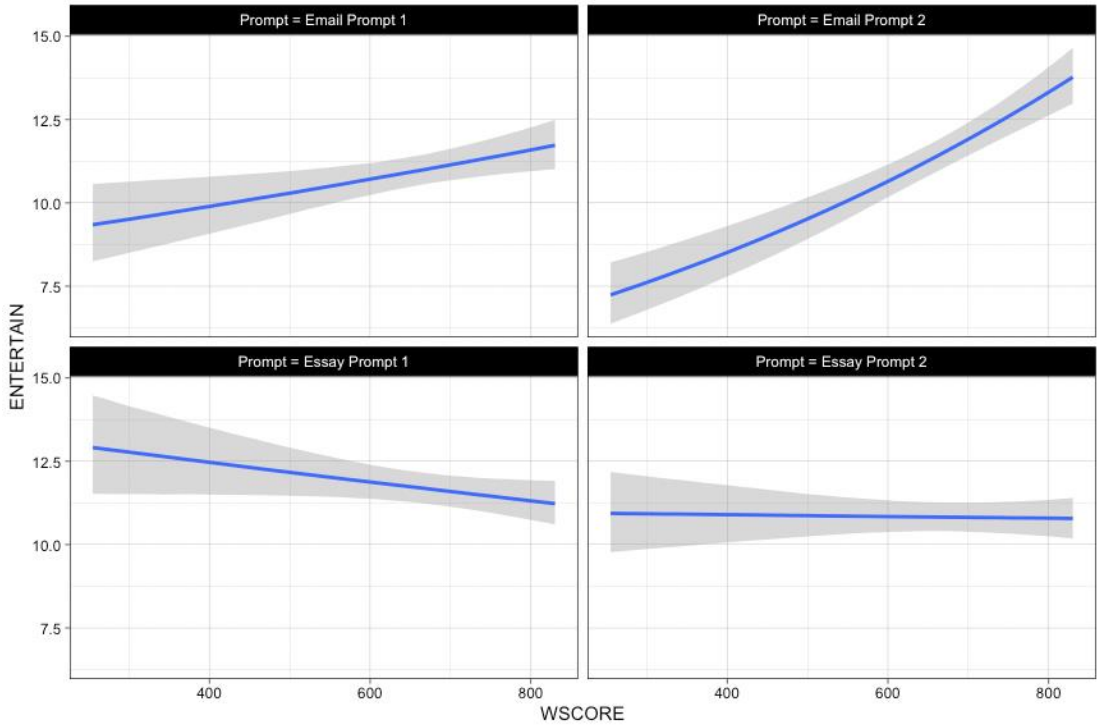


Figure 5.8

Slopes for writing score and JUSTIFYING strategy by prompt.

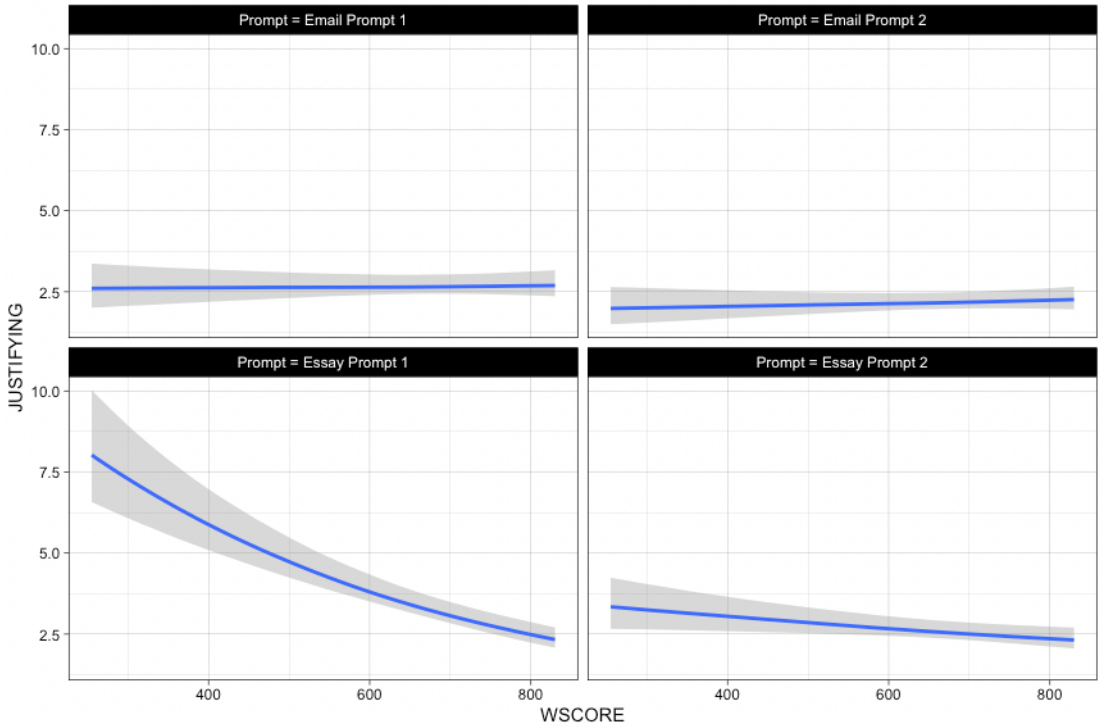


Figure 5.9.
Slopes for writing score and COUNTER strategy by prompt.

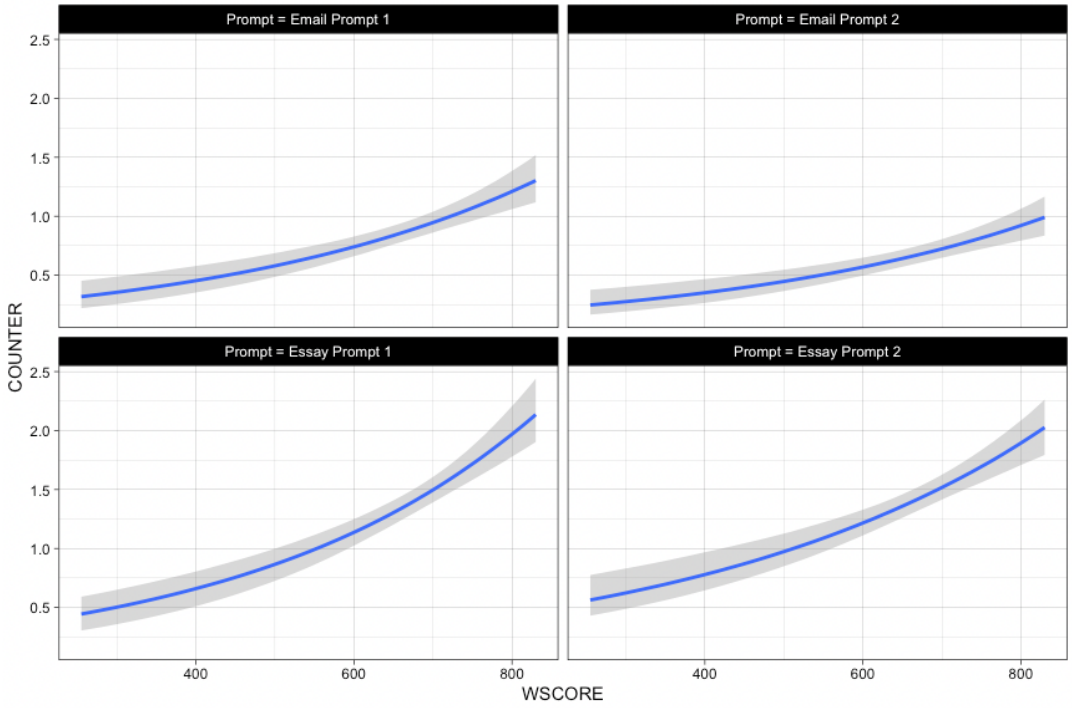


Figure 5.10
Slopes for writing score and DENY strategy by prompt.

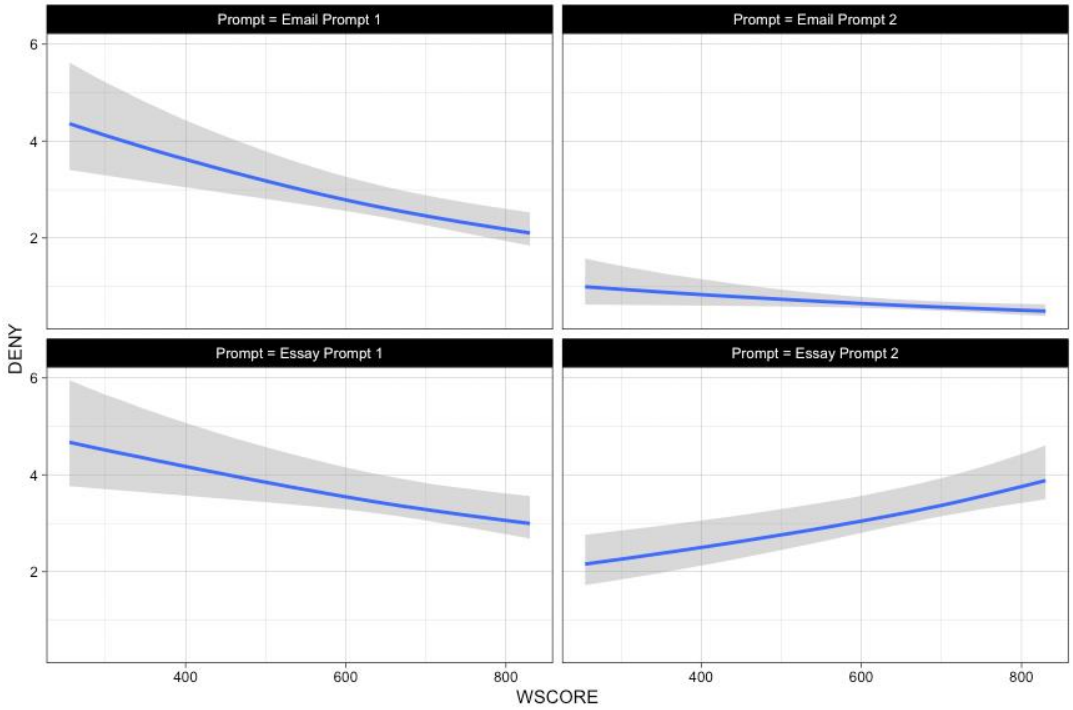
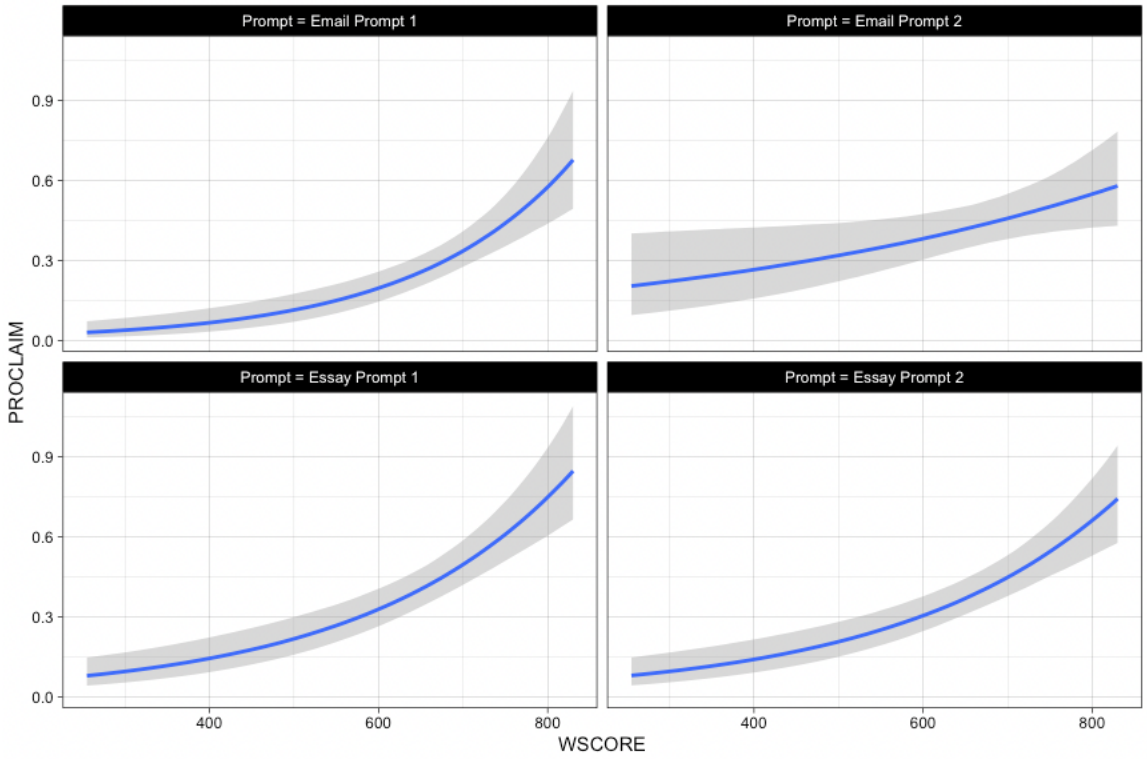


Figure 5.11

Slopes for writing score and PROCLAIM strategy by prompt.



5.4.4 RQ 3: To what extent does the use of Engagement strategies explain the writing score above and beyond the existing linguistic measures?

As a preliminary analysis for RQ3, bivariate correlation coefficients were examined between Writing scores, text lengths, frequency counts of Engagement strategies, and four rhetorical diversity measures. The results, presented in Table 5.10, indicate that none of the individual Engagement strategies correlate beyond the field-specific effect size benchmark of >.25 (Plonsky & Oswald, 2014). The correlations between Engagement strategies and Engagement diversity indices (gray areas in Table 5.10) show which of the Engagement strategies had more impact on each measure of rhetorical diversity. As expected from the formula (see methods), type I indices (i.e., Shannon's H') were more sensitive to picking up changes in

rare categories (e.g., ATTRIBUTION, COUNTER, DENY). In contrast, Type II indices, particularly Simpson's E, exhibit higher correlations with dominant categories (e.g., ENTERTAIN, MONOGLOSS). There is high collinearity between Shannon's H' and Simpson's D indices (rho = - .946). However, Shannon's H' shows a significantly higher correlation with writing score than Simpson's D index ($z = 9.7065$; $p < .001$ using a paired sample Fisher's r-to-z transformation and the *cocor* package), suggesting that Shannon's H' may capture more variance in the regression model. For this reason, Simpson's D index was dropped from the regression model to prevent any issues arising due to collinearity.

To investigate whether engagement measures explain above and beyond existing linguistic measures (RQ3), a total of four competing linear regression models were constructed for the following blocks of linguistic features—(a) text length, (b) lexical and phraseological features, (c) syntactic features, (d) cohesion features, and (e) all linguistic features. The relative advantages of adding Engagement features were assessed by comparing with the LOOIC index (Vehtari et al., 2022).

Table 5.10

Bivariate correlation coefficients (rho) between writing scores, numbers of words, and measures of Engagement strategies.

	WSCORE	Word Count	*ENT*[a]	*ATTR*[a]	*JUST*[a]	*MONO*[a]	*COUNT*[a]	*DENY*[a]	*PRCLM*[a]	*Shannon*	*Simpson D*	*Simpson E*
Word Count	**0.659**	1										
ENTERTAIN[a]	0.064	-0.025	1									
ATTRIBUTION[a]	0.165	0.130	-0.024	1								
JUSTIFYING[a]	-0.036	0.024	0.010	-0.001	1							
MONOGLOSS[a]	-0.169	-0.134	**-0.527**	-0.064	-0.063	1						
COUNTER[a]	0.232	0.191	0.017	0.186	-0.079	-0.165	1					
DENY[a]	0.038	0.105	0.050	0.205	0.124	-0.164	**0.272**	1				
PROCLAIM[a]	0.211	0.195	-0.025	0.058	-0.045	-0.160	0.052	-0.057	1			
Shannon's H'	**0.280**	**0.315**	**-0.351**	**0.579**	**0.230**	0.078	**0.502**	**0.505**	**0.263**	1		
Simpson's D	-0.159	-0.207	**0.522**	**-0.459**	**-0.295**	-0.243	**-0.404**	**-0.494**	-0.190	**-0.946**	1	
Simpson's E	**-0.305**	**-0.305**	**-0.645**	-0.177	0.183	**0.535**	-0.143	0.053	-0.181	0.092	**-0.375**	1

Note. Superscript [a] indicates normalized frequency per 100 words; Spearman correlation coefficients are used; Bold-face indicates pairs of correlation larger than a small size of effect (rho > .25) according to the field-specific effect size benchmark (Plonsky & Oswald, 2014).

5.4.4.1 Comparison with the number of words

To examine whether Engagement measures explain writing score after controlling for text length, four models were compared (Table 5.11): Text length only (A1), Engagement diversity (A2), Engagement diversity and text length (A3), and Engagement diversity, Normalized frequency counts of seven Engagement strategies (A4). Table 8 presents comparisons of LOOIC, WAIC, R2 conditional, R2 marginal, and R2 adjusted. Model weights based on LOOIC indicate that the Engagement diversity and text length model (A3) fits the available data best, accounting for 46% of the variance in writing score on adjusted R2. This is followed by two models: text length + Engagement diversity (A2) model and text length only (A1), each accounting for around 43% and 39% of variance. This result indicates that adding Engagement measures to the text length model improved the model's prediction.

Table 5.11
Summary of model comparison for text length and Engagement measures.

Model	LOOIC (weight)	LOOIC (SE)	WAIC (weight)	*R2 (cond.)*	*R2 (marg.)*	*R2 (adj.)*
A1) Text length only	6271.5 (<.001)	113.56	6271.5 (<.001)	0.39	0.39	0.39
A2)Text length + Engagement Diversity	6105.7 (0.20)	112.76	6105.5 (<.001)	0.43	0.43	0.42
A3) Text length + Engagement Diversity & Normed freq	**5960.1 (0.78)**	**105.51**	**5959.8 (>.999)**	**0.46**	**0.39**	**0.45**
A4) Engagement Diversity & Normed freq	6750.7 (0.02)	93.66	6750.5 (<.001)	0.28	0.29	0.26

Note. LOOIC stands for the Leave-One-Out cross-validation Information Criterion; SE stands for Standard Error; WAIC stands for the Watanabe-Akaike (Widely-Applicable) Information Criterion; cond. Stands for conditional; marg. Stands for marginal; adj. stands for adjusted.

5.4.4.2 Comparison with lexical and phraseological features

Another set of model comparisons was performed on lexical and phraseological features. Table 5.12 shows the results of this comparison. LOOIC showed that lexical and all Engagement indices (B3) showed the best fit to the data, explaining 44% of the variance in writing score. This is followed by the lexical and Engagement diversity features (B2; 42% of variance), and lexical feature only (B1) models. The differences between the lexical features only model (B1; 34% of variance) and lexical and Engagement features models (B2 and B3) are moderate with an approximately 10-point increase in the adjusted R^2 as well as a substantial decrease in LOOIC.

Table 5.12
Summary of models comparison for lexical and rhetorical features.

Model	LOOIC (weight)	LOOIC (SE)	WAIC (weight)	R2 (cond.)	R2 (marg.)	R2 (adj.)
B1) Lexical features only	6452.2 (0.06)	107.74	6452.1 (<.001)	0.36	0.37	0.34
B2) Lexical + Engagement Diversity	6107.1 (0.01)	99.82	6107.1 (<.001)	0.43	0.45	0.42
B3) Lexical + Engagement Diversity & Normed freq	**6011.9 (0.87)**	**99.72**	**6011.7 (>.999)**	**0.46**	**0.47**	**0.44**
A4) Engagement Diversity & Normed freq	6750.7 (0.06)	93.66	6750.5 (<.001)	0.28	0.29	0.26

Note. LOOIC stands for the Leave-One-Out cross-validation Information Criterion; SE stands for Standard Error; WAIC stands for the Watanabe-Akaike (Widely-Applicable) Information Criterion; cond. stands for conditional; marg. stands for marginal; adj. stands for adjusted.

5.4.4.3 Comparison with syntactic features

Another set of model comparisons was performed for syntactic measures (Table 5.13)—syntactic features only (C1), syntactic features and Engagement diversity (C2), syntactic features, Engagement diversity and individual strategies (C3), and Engagement measures only models (A4). The model comparison using LOOIC resulted in a similar picture to that of the Lexical features model. Syntactic features and Engagement diversity and individual strategies (C2 and C3) explain 33 and 35% of variance, respectively, while the syntax-only model (C1) explains 17% of variance.

Table 5.13
Summary of models comparison for syntactic and rhetorical features.

Model	LOOIC (weights)	LOOIC (SE)	WAIC (weights)	R2	R2 (marg.)	R2 (adj.)
C1) Syntax features only	7077.7 (0.04)	92.2	7077.7 (<.001)	0.18	0.19	0.17
C2) Syntax + Engagement Diversity	6508.2 (0.13)	94.5	6508.2 (<.001)	0.34	0.35	0.33
C3) Syntax + Engagement Diversity & Normed freq	**6410.0 (0.82)**	**93.73**	**6409.7 (>.999)**	**0.37**	**0.38**	**0.35**
A4) Engagement Diversity & Normed freq	6750.7 (3.00e-03)	93.66	6750.5 (<.001)	0.28	0.29	0.26

Note. LOOIC stands for the Leave-One-Out cross-validation Information Criterion; SE stands for Standard Error; WAIC stands for the Watanabe-Akaike (Widely-Applicable) Information Criterion; cond. stands for conditional; marg. stands for marginal; adj. stands for adjusted.

5.4.4.4 Comparison with cohesion measures

The pattern of results for cohesion features was similar to that for lexical and syntactic features (Table 5.14). That is to say, the model with cohesion features, Engagement diversity, and individual Engagement strategies (D3) was the most predictive for writing scores, with an additional 16% of variance explained by cohesion features alone (D1).

Table 5.14
Summary of model comparison for cohesion and rhetorical features.

Model	LOOIC (weight)	LOOIC (SE)	WAIC (weight)	R2	R2 (marg.)	R2 (adj.)
D1) Cohesion features only	6966.4 (0.04)	91.06	6966.4 (<.001)	0.22	0.23	0.2
D2) Cohesion + Engagement Diversity	6556.3 (0.07)	93.77	6556.2 (<.001)	0.33	0.34	0.32
D3) Cohesion + Engagement Diversity & Normed freq	**6395.6 (0.89)**	**96.18**	**6395.4 (>.999)**	**0.38**	**0.38**	**0.36**
A4) Engagement Diversity & Normed freq	6750.7 (<.001)	93.66	6750.5 (<.001)	0.28	0.29	0.26

Note. LOOIC stands for the Leave-One-Out cross-validation Information Criterion; SE stands for Standard Error; WAIC stands for the Watanabe-Akaike (Widely-Applicable) Information Criterion; cond. stands for conditional; marg. stands for marginal; adj. stands for adjusted.

5.4.4.5 Engagement features on top of all other linguistic features

Finally, I was interested in whether Engagement measures would be predictive of writing score above and beyond all other linguistic measures at the levels of lexis, syntax, and cohesion. To this end, I constructed four models and compared model performance (Table 5.15). The results mirrored those of individual comparisons with each of the lexical, syntactic, and cohesion

features. In other words, the addition of rhetorical features of Engagement (the model E3) contributed to model prediction, adding almost 6% of additional variance on top of the combination of all other features (E1). This was followed by Linguistic measure + Engagement diversity (E2; 49% of variance explained). When the linguistic features model (E1) was compared to the rhetorical features only model (A4), the former demonstrated a better fit to the data, suggesting that rhetorical features do not replace existing measures but rather complement extant measures. Overall, the results of the model comparisons strongly support the inclusion of rhetorical features of Engagement strategies in predicting writing scores.

Table 5.15
Summary of model comparisons between all linguistic measures and rhetorical features.

Model	LOOIC (weights)	LOOIC (SE)	WAIC (weights)	R2 (cond.)	R2 (marg.)	R2 (adj.)
E1) Linguistic features only	5944.7 (0.11)	99.17	5944.5 (<.001)	0.48	0.49	0.45
E2) Linguistic + Engagement Diversity	5763.7 (<.001)	97.92	5763.4 (<.001)	0.51	0.52	0.49
E3) Linguistic + Engagement Diversity & Normed freq	**5678.3 (0.85)**	**98.65**	**5677.8 (>.999)**	**0.53**	**0.53**	**0.51**
A4) Engagement Diversity & Normed freq	6750.7 (0.04)	93.66	6750.5 (<.001)	0.28	0.29	0.26

Note. LOOIC stands for the Leave-One-Out cross-validation Information Criterion; SE stands for Standard Error; WAIC stands for the Watanabe-Akaike (Widely-Applicable) Information Criterion; cond. stands for conditional; marg. stands for marginal; adj. stands for adjusted.

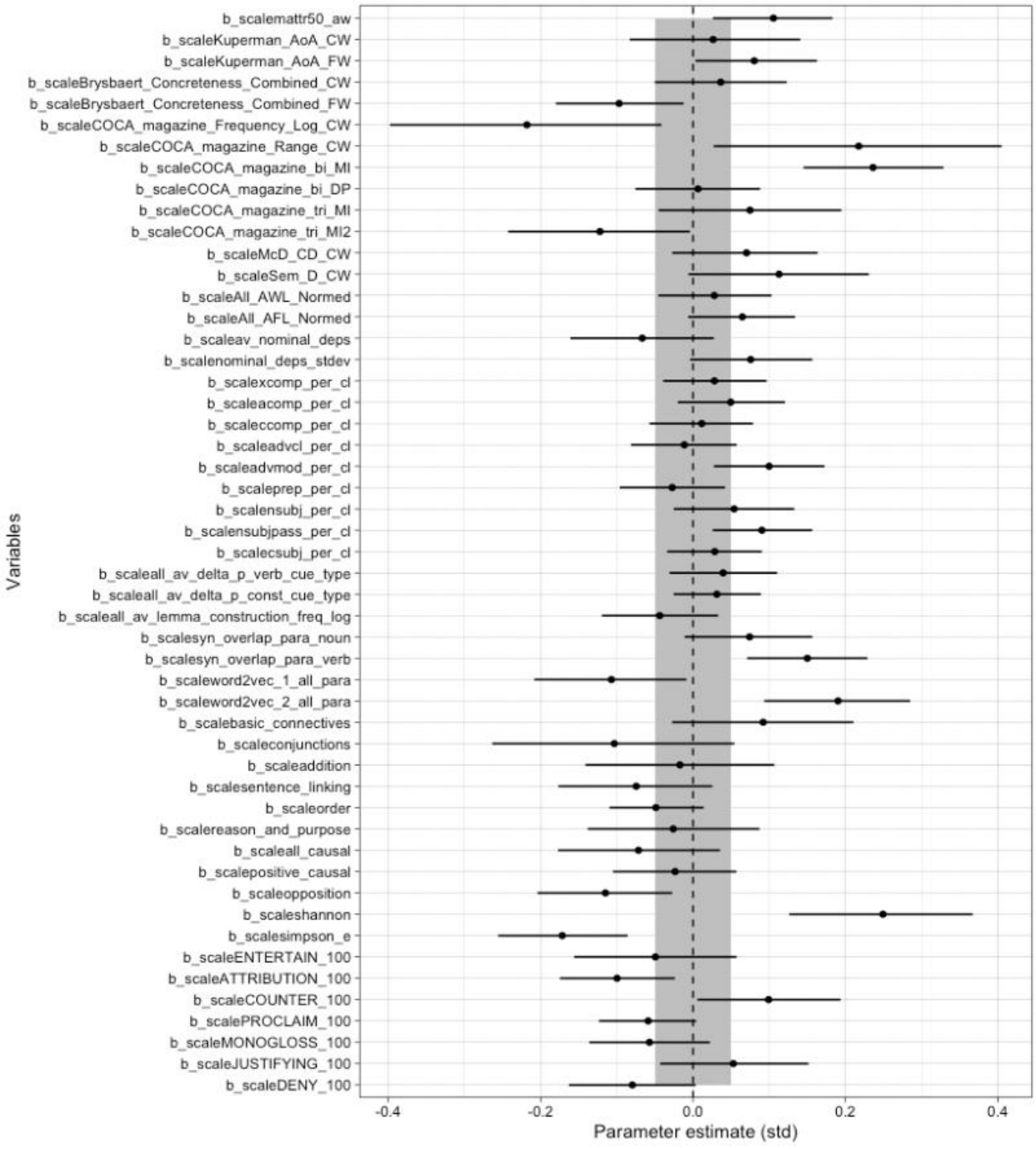
Figure 5.12 summarizes the parameter estimates of the best model from Table 5.15 (Linguistic + Engagement Diversity & Normed Freq). Although direct comparisons of standardized parameter estimates are not recommended (Mizumoto, 2022), Figure 5.12 can suggest which variables tend to impact on prediction. Predictors which do not include the Region Of Practical Equivalence (ROPE; e.g., Kruschke, 2014; Vasishth & Gelman, 2021) in their 95% Credible Intervals were the following five indices: (1) COCA magazine Bigram MI

(Phraseological sophistication), (2) Synonym Overlap Verb (Cohesion), (3) Word2Vec All Paragraph 2 (Cohesion), (4) Shannon's H' index (Rhetorical), and (6) Simpson's E index (Rhetorical). Therefore, the final model contained several predictors from each linguistic domain, except syntax. Notably, two of the Engagement diversity indices (Shannon's H' and Simpson's Evenness) were also predictive in the final prediction model.

5.5 Discussion

The present study aimed to explore how the rhetorical features of Engagement relate to writing performance across two task types—email and argumentative essay writing. It also examined the extent to which a new set of Engagement measures explain assessed writing scores above and beyond existing linguistic features at the levels of lexis, syntax, and cohesion. The results of Poisson regression analyses (RQ 1 and 2) indicated that two task types may elicit distinct patterns of rhetorical features, particularly with regard to the number of MONOGLOSS statements and hence more heteroglossic strategies (e.g., ATTRIBUTION, COUNTER). The results also show trends that were common across the two task types. These common trends include a general decrease in MONOGLOSS as a function of writing scores, and increases in ATTRIBUTION, COUNTER, and PROCLAIM. The results of linear mixed-effects regression and model comparisons (RQ3) revealed the key role that the rhetorical features of engagement strategies play in predicting writing scores above and beyond existing linguistic features. In the remaining sections of this chapter, I will briefly discuss these results.

Figure 5.12
Posterior distributions of parameters on a standardized scale.



Note. The gray area indicates the Region Of Practical Equivalence (ROPE), which is in the range -0.05 to 0.05 for the current model (cf. Kruschke, 2014).

5.5.1 RQ1: Task type comparisons—Argumentative essay prompts may elicit more heteroglossic statements

The results of Poisson regressions revealed that heteroglossic statements were more frequently observed in argumentative essay prompts. This was contrary to the theoretical prediction based on the situational requirements of these two distinct settings of writing. Specifically, the current study hypothesized that email prompts would elicit more heteroglossic strategies than essay writing because an email prompt can specify a clear, albeit imaginary, recipient, who supposedly has more power over the test-taker (i.e., city mayor or new school principal). This clearly imagined audience of the task was initially expected to set the tenor of the discourse more clearly (Halliday & Matthiessen, 2014), making interpersonal language use more important, thus leading to more frequent use of heteroglossic strategies. Although the results showing divergent distributions of Engagement strategies across task types do indeed suggest that the contextual parameters of the writing task may alter the test-takers' use of rhetorical strategies, the direction of the effect was the opposite from the theoretical prediction. This opposite pattern from the prediction requires a more detailed explanation, which is addressed below.

Although speculative, a few possible reasons can explain the current results. The first possible explanation concerns the fact that the argumentative genre invoked the use of interpersonal language without setting a clear audience. The result of more frequent heteroglossic strategies in argumentative essay tasks can be taken to indicate that even without a specified audience in mind, the writer employs a range of rhetorical devices to convince potential readers of their writing. In the case of standardized English tests, the default strategy may be that the test-taker convinces the raters. Therefore, although implicit, the nature of the argumentative

writing task may allow the writer to have some addressee in mind. Relatedly, the same result can be interpreted in relation to the ubiquitous nature of evaluative language (Halliday & Matthiessen, 2014; Hunston & Thompson, 2000; Martin & White, 2005; Thompson & Alba-Juez, 2014). On this account, any instance of language use to some extent invokes interpersonal language use (Hunston & Thompson, 2000; Martin & White, 2005). In other words, argumentative essays are able to elicit engagement strategies because the overall social purpose of argumentation is fundamentally interpersonal (i.e., convincing putative readers). In light of this finding, the use of an argumentative essay as a task type might be justified as a generic task type, despite its apparent lack of specific contextual parameters (cf. Schneer, 2014) and thereby situational authenticity (Bachman & Palmer, 1996, 2010). However, more research is needed to justify the use of the traditional argumentative genre in relation to variety of genre types (e.g., critique, discussion) because the findings indeed suggest that the two task types differ in the frequency of engagement strategies.

A second possible explanation is that genre-specific strategies in email tasks, such as salutations (e.g., My name *is* XXXX; I *am* a high school student at…) and closing (e.g., I *hope* that you take these points into consideration), contributed to the general increase in MONOGLOSS in email tasks. While this interpretation may not necessarily pose a validity question regarding task as an elicitation tool in standardized writing exams, it highlights the relatively coarse-grained nature of the MONOGLOSS category in engagement system (S. H. Lee, 2017). In the literature on Engagement, there is an ongoing discussion on whether the MONOGLOSS category is granular enough to capture different facets of argumentative discourse. Specifically, researchers have argued for the necessity of distinguishing factual statements from bare assertions (e.g., S. H. Lee, 2017). It seems that Martin & White (2005)

presuppose that appraisal analysis is applied to an argumentative discourse where obvious factual statements are not the focal point of the analysis. However, as Lee (2017) demonstrates, MONOGLOSS strategies can be subdivided into several types, some of which are not particularly relevant in a discussion of stance expressions. Therefore, this finding is in line with Lee's argument that more granular distinction will enhance the interpretability of the MONOGLOSS category. At the same time, it is also important to remind ourselves that Engagement is only one of the dimensions of evaluative language proposed by Martin & White (2010), who focus on the epistemic dimension of discourse. More emic classifications of rhetorical strategy may still be needed to capture nuanced language use. It may be the case that MONOGLOSS statements involved attitudinal stances (e.g., It is important; I am impressed, disappointed, etc.; Biber & Finegan, 1989). Thus, future research should clarify how different aspects of evaluative language are used to characterize email writing tasks.

Despite the general opposite patterns, a close look at patterns of ENTERTAIN and COUNTER may still be interpreted in line with expected genre effects. Although ENTERTAIN was the most frequent strategy in both task types, it tended to predict writing scores more strongly in Email writing than in Essay writing tasks. In other words, the additional ENTERTAIN strategies contributed to higher writing scores on the Email prompt, but not the argumentative essay. This pattern is in accordance with the initial hypothesis that interpersonal language use is more important in email writing tasks than argumentative ones. Similarly, although more COUNTER strategies are indicative of higher writing scores, this strategy was less frequent in the email writing task than in the argumentative essay (see Figure 5.4). Based on this hypothesis, it is speculated that the clear tenor of the discourse set by the email writing task resulted in less use of a disclaim strategy. When disclaim strategies (i.e., DENY and COUNTER)

are used, the writer essentially takes two positions and rejects one of them to support their claim. Such a rejection of an alternative position may raise interpersonal demands; hence, the two options may have been dispreferred in the email task, where the test-taker was expected to write to someone who possesses more power relative to them. All in all, although the hypothesis that email writing elicits more diverse heteroglossic strategies was not confirmed, some of the strategies were partially in line with the theoretical prediction. In future research, more detailed, specific hypotheses need to be established regarding the one-to-one mapping of contextual parameters and individual engagement strategies.

5.5.2 RQ2: Engagement strategies and writing score—A high writing score is associated with more COUNTER and PROCLAIM and decreased MONOGLOSS.

The results of the Poisson regression analyses indicated that higher writing scores were associated with decreases in MONOGLOSS strategies but increases in COUNTER and PROCLAIM regardless of the prompt selection. The frequency of other strategies also tended to increase as a function of writing scores, such as ENTERTAIN in email tasks and ATTRIBUTE in argumentative essay tasks. Such a shift from MONOGLOSS-dominant to more diversified heteroglossic strategies has been well documented in previous literature on engagement, which investigated the qualities of students' disciplinary essays (e.g., Lancaster, 2014; Wu, 2007). In particular, increases in COUNTER strategies have been repeatedly observed in the literature (e.g., Lancaster, 2014; Wu, 2007). In this study, although the effect size was not large, a 1SD increase in writing scores corresponded to 1.24 times the frequency of COUNTER strategies and 1.39 times the frequency of PROCLAIM, on average. This result indicates that the relative distribution of rhetorical features of Engagement is not only a characteristic of disciplinary-specific writing but also a feature of second-language timed writing tasks. This finding strongly

aligns with the theoretical underpinnings of the engagement system (Martin & White, 2005) and, by extension, with the notion of evaluative language (Hunston & Thompson, 2000; Xie, 2020), which emphasizes the ubiquitous nature of interpersonal language features. Therefore, the overall results support the importance of assessing the interpersonal dimensions of language use, even in standardized language testing situations.

Furthermore, the fact that high-scoring writing in this study was closer to highly evaluated disciplinary writing in terms of Engagement can be taken to indicate that L2 timed essays may still be able to tap into some aspects of the rhetorical features which are important in disciplinary writing. More research is needed to clarify the nature of timed essay writing tasks in standardized English assessments in terms of their interpersonal language features.

5.5.3 RQ3: Prediction model of writing score—Engagement strategies related to writing scores above and beyond existing linguistic measures

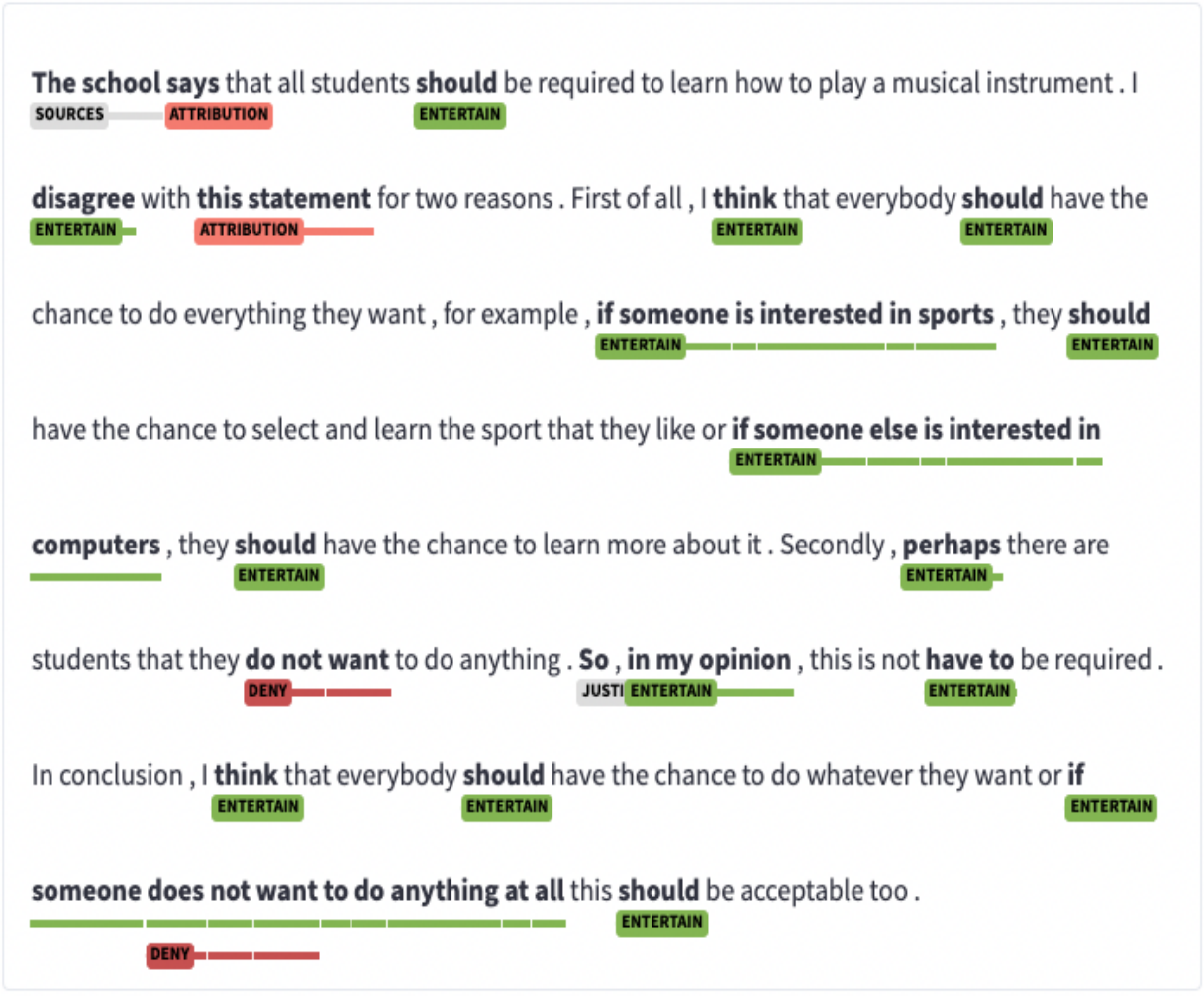
A series of regression analyses and model comparisons (RQ3) indicate that the rhetorical features of Engagement were able to explain writing scores above and beyond existing linguistic measures at the levels of lexis, phraseology, syntax, and cohesion (Tables 5.8–12). The final model comparison presented in Table 5.12 between all linguistic indices (E1) versus the Engagement indices-only model (A4) also indicated that the rhetorical features of engagement are not a replacement for existing linguistic measures, but they work in harmony when predicting writing scores (as demonstrated in model E3). This result can be interpreted as evidence of the advantage of including rhetorical features in Automated Writing Evaluation (AWE) engines.

Figures 5.13 and 5.14 present the output of the Engagement Analyzer for two argumentative essay writing tasks with varying engagement diversity and essay scores. The first

essay (Figure 5.13) has a lower Shannon's H' index (H' = 1.4; lower than the 20th percentile) and the second essay (Figure 5.14) is high in diversity measures (H' = 2.47; around the 75th percentile). The first essay is characterized by dominance of ENTERTAIN strategies (modal verbs *should*, if-clause for conditional statement) with occasional ATTRIBUTION (the school *says*) and DENY (*do not want*). The effect of this is that it consistently takes a particular stance while recognizing other possible stances. This tone of argument is consistent throughout. On the other hand, the second essay uses both Expansion and Contraction strategies to negotiate the floor. It does not only use Expansion strategies such as ENTERTAIN with differing degrees of certainties (*may*, *might*, *firmly believe*) but also uses a COUNTER strategy twice to present changes of tone in the argument. The first COUNTER strategy appeared to be used as a rhetorical device to expand the discussion (COUNTERing their own introduction, which states the writer's stance on the discussion). The second COUNTER towards the end (i.e., *but*) is used to reiterate the main points of the argument and their essential stance. These alternations between expansion and contraction strategies have been identified as one distinguishing feature of high-scoring essays in previous study (e.g., Lancaster, 2014). It is also noteworthy that the PROCLAIM strategies in the second example seem to include both CONCUR-concede and PRONOUNCE subtypes, which are not implemented in the current version of the Engagement Analyzer. Thus, if the Engagement Analyzer is accurate enough to conduct a finer-grained analysis, it might be able to reveal more nuanced rhetorical strategies in the second example. In summary, the findings of the current study demonstrate that the diversity of rhetorical features of Engagement explains aspects of discourse in argumentation not already explained by the individual features of lexis, syntax, and cohesion.

Figure 5.13

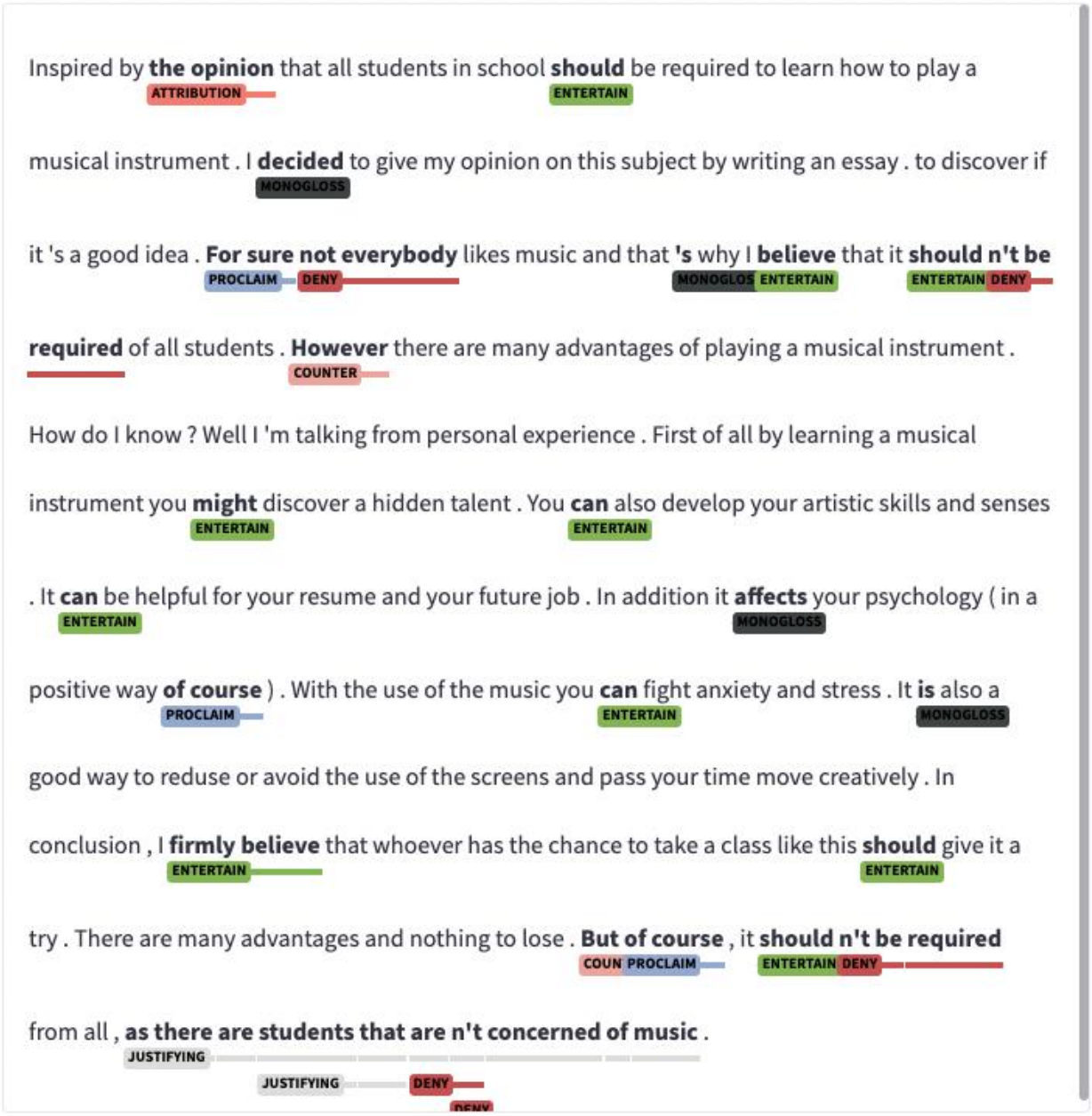
Example test script with lower Engagement diversity.



Note. Shannon's H' = 1.41; 20th percentile; ID = 50210110

Figure 5.14

Example test script with higher Engagement diversity.



Note. Shannon's H' = 2.47; ID = 40204894.

5.5.4 Implications for practice and research on language assessment

The findings of this study have several implications for the practice of second language assessment. First, the results of the regression analyses and model comparisons (RQ3) show the benefits of taking rhetorical features of writing into consideration in automated scoring models. Because existing automated scoring systems exclusively focus on lexical and grammatical information in essay responses (e.g., Attali, 2007; Enright & Quinlan, 2010), including rhetorical features would increase the construct coverage of assessment tools. This would enhance some important aspects of assessment use argument, including explanation and extrapolation inferences (Chapelle et al., 2008), as well as the meaningfulness of the assessment record (Bachman & Palmer, 2010). Specifically, the inclusion of rhetorical features probably enhances the ability of test scores to reflect test-takers' ability to perform real-world academic tasks (i.e., term papers), reducing the misclassification of students' ability in matriculating them to different levels of English for Academic Purposes (EAP) courses, for instance. Although this study can only be taken as an initial indication of the benefits of rhetorical features in an automated assessment model, this agenda should receive more attention in future studies to improve the classification accuracy of AWE systems, and the construct definitions of writing skills more broadly.

Second, the findings of the current study have implications for task designs in writing assessments. Although the patterns of rhetorical features elicited in email and argumentative were the opposite of what was predicted, this fact implies that our understanding of task type and resulting linguistic use needs to be carefully scrutinized in future studies. In line with this argument, various research has already been conducted to reveal differences in linguistic production across task types (e.g., Alexopoulou et al., 2017; Michel et al., 2019). While this

research has focused on lexical and grammatical complexity, no research has focused on rhetorical features across task types. More research is needed to reveal the exact effects of task manipulation on linguistic production at the levels of discourse functions, including the rhetorical features of engagement.

Research on interpersonal language use likely receives more attention in the language assessment with the recent announcement from the Educational Testing Service (ETS) about the addition of Writing for an Academic Discussion task to TOEFL iBT®. According to ETS, the old independent writing task will be replaced by the academic discussion task, which requires the test-taker to read a question from the professor as well as responses from the imagined classmates and then respond to the online discussion (https://www.ets.org/toefl/transcript/writing-for-an-academic-discussion-task.html). After the new format of the TOEFL iBT is launched on July 26th, 2023, the students' writing response may include different types of interpersonal language than would have been elicited by the traditional independent essay prompt. However, the new rubrics for this Writing for an Academic Discussion task (as of April. 24th, 2023) appear to lack the criteria for interpersonal language use or pragmatic features of the responses (https://www.ets.org/pdfs/toefl/toefl-ibt-writing-rubrics-enhanced.pdf). Thus, it is expected that more research will be conducted on how to best assess interpersonal language use in a large-scale assessment. As TOEFL iBT has been using Automated Writing Evaluation (AWE) engine as one of the raters in the rating sessions, the modification in the language elicitation task will likely result in a revision of the AWE systems in the near future.

5.5.5 Limitations

A few limitations are worth noting. First, although interclass correlation analysis showed the internal consistency of Engagement measures across various versions of the Engagement Analyzer (ICC > .8), there is still room for improving these models' item-level performance. However, as in Study 2, I employed the same four separately trained NLP pipelines and statistically controlled for their variabilities through the mixed-effect models. To increase the model's accuracy on item-level annotation, it would be beneficial to increase the size of the manually annotated corpus.

Second, although the current study has demonstrated a clear benefit of applying indices of heterogeneity and evenness often used to measure biodiversity, the reliability of these measures against text length are not fully addressed in the present study. It is still possible for the Shannon's and Simpson's indices used in this study to be *inherently* related to text length. In second language research, researchers pay close attention to the issue of the text length stability of lexical variety measures (e.g., Koizumi & In'nami, 2012; Zenker & Kyle, 2021). Future studies should test whether rhetorical diversity measures are insensitive to text length.

Third, the between-participant nature of the current task type comparison limits the generalizability of its findings. Since no individuals completed both task types, it was not possible to partition the variances due to task type and individuals. More research with a rigorous within-participant design is needed to clarify task type differences in the use of rhetorical features.

5.6 Chapter Conclusion

This chapter has presented a study investigating the relationship between the rhetorical features of Engagement and assessed essay quality in two task types—email and argumentative essay tasks. It has demonstrated that the inclusion of rhetorical features measured in the Engagement Analyzer can explain the variance in writing scores above and beyond existing linguistic measures at the levels of vocabulary, phraseology, syntax, and cohesion. The model comparisons indicated that at least additional 6% of the variance may be attributed to the engagement strategies and their diversity in the writing. The findings of the current study illustrate the complementary nature of existing linguistic features and new measures of rhetorical features when describing more or less successful timed essays.

CHAPTER 6 CONCLUSION

6.1 Chapter overview

In the previous three chapters, I presented three empirical studies that aimed to (a) develop and evaluate a new natural language processing (NLP) tool to annotate engagement resources (Martin & White, 2005), (b) apply the new tool to describe the stance-taking features of written university assignments across different registers, and (c) use the measures of stance-taking features to predict writing proficiency in large-scale standardized L2 English proficiency tests. In this chapter, I briefly summarize the findings of the three studies and outline the overall contribution of this dissertation research. Finally, I conclude the dissertation by calling for large-scale collaboration between discourse analysts, applied linguists, and computational linguists to produce high-quality linguistic annotations that allow the investigation of important textual features in a highly replicable manner.

6.2 Summary of Findings

6.2.1 Study 1: Engagement Analyzer

In the first study, I aimed to develop an end-to-end natural language processing system that can undertake engagement resource analysis, drawing on Appraisal framework (Martin & White, 2005). To this end, I compiled a gold-standard annotated dataset, the Engagement Discourse Treebank (EDT), to train and evaluate a machine-learning system. With the help of two trained annotators, 126,411 tokens (4,688 sentences) were manually annotated and thoroughly reviewed by the researcher, comprising a gold-standard dataset for Engagement resource analysis. After ten weeks of training sessions and iterative refinement of annotation guidelines (Fuoli, 2018), the intercoder agreement between trained annotators was estimated to

be moderate (Cohen's Kappa = .67). The intercoder agreement also highlighted several challenging categories to annotate: PROCLAIM (F1 = .4), SOURCES (F1 = .57), ATTRIBUTION (F1 = .6), and ENDOPHORIC (F1 = .62). These figures were considerably higher than that in a previous annotation project by Read and Carroll (2012). In Chapter 3, I discussed the challenging nature of these categories in relation to previously reported attempts to make sense of Martin & White's (2005) discourse framework in real-world corpus annotation (Fuoli, 2018). The chapter concluded that a more transparent corpus annotation project like this one would facilitate collaborations between discourse analysts, corpus linguists, and computational linguists in order to develop a more rigorous annotation scheme for a given discourse-oriented annotation task.

The remaining parts of Study 1 reported a machine-learning experiment where the current version of Engagement Discourse Treebank (EDT) was used to train the span categorizer component of spaCy (Honnibal et al., 2014/2020). An experiment with a total of three different neural architectures and four different versions of pre-trained (and domain-adapted) RoBERTa models (Liu et al., 2019) revealed that the neural network model could perform as well as, or even outperform, the human baseline (macro F1 = .728 averaged over 5-fold Cross-Validation). Notably, the results indicated that while the ML system tended to struggle with tags that human coders struggled with, it was considerably better than human coders at producing consistent tags with a gold standard. The study concluded with future directions that pertain to developing applied-linguistics-oriented NLP tasks.

6.2.2 Study 2: Engagement strategies across university registers

The goal of Study 2 was to describe the registers of university written assignments from an engagement perspective. To this end, four separately trained versions of the Engagement Analyzer were used to analyze 2,685 single-genre-family assignments from the British Academic Written English corpus (Alsop & Nesi, 2009). The goal of the analysis was to reveal what combinations of assignment-related or writer-related factors explain the distribution of Engagement strategies. The results of a Bayesian multilevel MANOVA indicated that two grouping factors tended to explain a large amount of variance in the distribution of Engagement strategies across categories—(a) two-way interaction between genre family and discipline and (b) individual writers. In contrast, other writer-related factors, including their L1 and secondary education, contributed little to determining Engagement strategies. The current findings suggest a nuanced picture of disciplinary and/or genre-based writing because the rhetorical features of Engagement will not be determined independently by discipline or genre family, but their specific combination matters. In addition, a large amount of variance was due to individual writers, which suggests that some individual "styles" (see Biber & Conrad, 2019) play a role when explaining the distribution of Engagement strategies. Finally, unlike in previous case studies (e.g., Lancaster, 2014; Wu, 2007), the current analysis did not find significant effects of assignment-related variables, such as levels or grades. I discussed possible method effects between the current study and previous research and recommended replicating the current findings with corpora that include lower-graded essays in future research. With the help of the Engagement Analyzer, however, such replication studies will be less labor-intensive, as the researchers can conduct (semi-)automatic Engagement resource analysis with reasonable accuracy. This point is discussed further in the statement of contribution section.

6.2.3 Study 3: Engagement strategies in timed L2 essays and their relation to essay quality

Study 3 sought to explore the interface between automated writing evaluation and discourse NLP by using measures derived from automated Engagement resource analysis to predict assessed essay quality alongside existing linguistic measures at the levels of lexis, syntax, and cohesion. To this end, two sets of operationalization of Engagement strategies were examined—normed frequencies of individual categories and diversity of Engagement strategies. In particular, the diversity of Engagement strategies was conceptualized based on the previously reported tendency for higher-rated disciplinary writing to use a wider variety of Engagement strategies (e.g., Lancaster, 2014; Wu, 2007). It was hypothesized that these diversity measures of Engagement can capture substantial variance in human ratings of e-mail and essay writing, above and beyond other linguistic measures. The results of a series of regression analyses indicated that Engagement diversity measures do indeed contribute to a prediction model of essay scores above and beyond existing linguistic measures. It was also found that the distribution of individual Engagement strategies may differ across two task types—e-mail or argumentative essay prompts. Although some across-task distributions of Engagement strategies were unexpected, several category-specific distributions aligned with predictions, including steeper slopes for ENTERTAIN on writing scores in e-mail prompts. Overall, the current study has demonstrated the potential uses of rhetorical features of Engagement in automated writing evaluation (Carr, 2013; Lu, 2021). That is to say, while the rhetorical features of Engagement will NOT replace existing linguistic measures at the levels of lexis, syntax, and cohesion, they can be used as supplementary features in automated essay scoring and evaluation systems to enhance Explanation and Extrapolation inferences (Attali, 2007; Chapelle et al., 2008; Enright & Quinlan, 2010).

6.3 Statement of Contribution

There are several ways in which the current study contributes to the literature on Appraisal analysis, natural language processing for applied linguistics, genre-based pedagogy in English for Academic Purposes, and automated writing evaluation.

6.3.1 Automated Engagement resource analysis

This study contributes to the literature on Appraisal analysis (Martin & White, 2005), and Systemic Functional discourse analysis more broadly (Hood, 2010; Martin & Rose, 2007). The contribution to this area of research takes two forms—the Engagement Discourse Treebank (EDT) and the Engagement Analyzer. As pointed out in the Literature Review (Chapter 2), extant research has shown the potential benefits of investigating rhetorical features of engagement (Lancaster, 2014; Wu, 2007; Xu & Nesi, 2019), but researchers have tended to take a small-scale case study approach, with or without quantification. The Literature Review suggested that this may primarily be due to the intensive nature of manual discourse analysis. Although one published study proposed a standardized appraisal annotation procedure (Fuoli, 2018), to the best of my knowledge, no subsequent studies have incorporated this procedure in their analyses of appraisal resources. Therefore, in developing the EDT and the Engagement Analyzer (Chapter 3), the current dissertation study has followed the principle of a stepwise annotation procedure (Fuoli, 2018). The results indicate that, with careful manual annotation, automatic analyses of rhetorical features of engagement are possible. This finding is promising in that the Engagement Analyzer can be used to reduce some of the costs in the analysis of discourse features, thus contributing to the scalability of analysis. Hopefully, this methodological development will serve as a basis for increasing replication and extension of research using an engagement system (e.g.,

Chang & Schleppegrell, 2011; Lam & Crosthwaite, 2018; Lancaster, 2014; Schad, Nicenboim, et al., 2021; Wu, 2007; Xu & Nesi, 2019).

6.3.2 Natural language processing for applied linguistics

The findings showing the capabilities of state-of-the-art NLP models to disambiguate rhetorical discourse categories have implications for the role of NLP in applied linguistics research in general. In this regard, the current study can be taken as a case study for developing an end-to-end machine learning model for the linguistic annotation of less commonly investigated discourse constructs. Given the promising findings in Study 1, the methodological procedure of the study can be used to develop an automated linguistic annotators with increasingly broader scope. To this end, extant research has already paid attention to features of metadiscourse (Hyland, 2005a), moves and steps (J. Swales, 2004; see Cotos & Pendar, 2015), and verb-argument structure constructions (Kyle & Sung, 2023). Ultimately, increasing numbers of automated linguistic annotation models will contribute to the research and practice of second language teaching and assessment in various contexts by providing highly efficient means to generating impromptu feedback on learner performance and more precise measurement of linguistic constructs in Automated Writing Evaluation systems (Attali, 2007; Enright & Quinlan, 2010). These three areas of application are extended in the following sections.

6.3.3 Corpus-based Register Analysis

The automated annotation of the functional category of stance-taking expressions has several implications for corpus-based register analyses (e.g., Biber, 1988, 2006a; Biber & Conrad, 2019). In this research tradition, researchers have investigated variations in lexico-

grammatical features across different situational variables, such as disciplines and genres, as examined in Chapter 4. Methodologically, researchers first describe the situational variables that interest them and then investigate how linguistic features vary across these situational contexts. In doing so, they often run factor analyses on many lexico-grammatical features to reveal "constellations" of co-occurring linguistic features in texts across corpora. Each constellation of lexico-grammatical features (or dimensions in Biber's terminology) is interpreted in relation to the possible functional situational motivations underlying their co-occurrences (Biber, 1988). This corpus-based framework, known as multidimensional analysis (MD or MDA), has been a dominant method in corpus-based register analysis since Biber's (1988) study on variations across spoken and written language (Biber, 2006a; Biber, Conrad, Reppen, et al., 2004; Kyle, Choe, et al., 2021; for a methodological synthesis see Goulart & Wood, 2021)

The automated engagement analysis presented in the current dissertation is highly compatible with corpus-based register analysis in its conception and methodology. That is to say, both approaches attempt to describe and understand the ways in which language production varies across registers and interpret them with respect to situational variables. More precisely, as demonstrated in Chapter 4, engagement strategies can be considered as dimensions to investigate register variations. Here, I characterize engagement strategies as "dimensions", treating them at the same level as MDA dimensions. This is because engagement strategies are interpreted at the level of discourse semantics in both theory and their operationalization. Using a probabilistic NLP approach, engagement analysis directly captures the functional communicative dimension without making inferences from individual lexico-grammatical "features". In other words, the current end-to-end NLP approach to engagement analysis attempts to model the function and can be reverse-engineered for "constellations" of lexico-grammatical features, if desired. As such, the

features of engagement strategies are arguably theoretically motivated "dimensions" in the sense that MDA dimensions are empirically derived (when created through Exploratory Factor Analysis). Future research may benefit from investigating additional interpersonal language features of engagement along with other MDA dimensions, such as Oral versus Literate Discourse, Procedural versus Content-focused discourse, Reconstructed accounts of events, and Teacher-centered stance (Biber, Conrad, Reppen, et al., 2004) to gain fuller insights into how registers vary in terms of these communicative functional dimensions.

6.3.4 Genre-based Pedagogy for English for Academic Purposes

As mentioned in Chapter 4, the Engagement Analyzer and its visualization tool have important implication to the genre-based pedagogy (the demo version freely accessible through HuggingFace Space at https://huggingface.co/spaces/egumasa/engagement-analyzer-demo)—It allows in-depth analyses and visualizations of user-input texts under the engagement framework. These features are useful for instructors and learners in materials development, feedback provision, and classroom activities. Research on Instructed Second Language Acquisition (ISLA) and genre-based pedagogy is generally in support of awareness-raising activities through textual enhancement or guided discussion of textual features (for a review see Boers, 2021; De Oliveira & Schleppegrell, 2015). The systemic functional genre-based pedagogy emphasizes the process of deconstruction, joint construction, and individual construction of text in writing instruction (Rose & Martin, 2012). The capability for instant visualizations enables one to apply the automated engagement analysis in any of these stages of writing pedagogy. For example, the instructor can input a model text from the target genre into the Engagement Analyzer and present the sequences and patterns of engagement to learners following the joint text-mining activity (as

part of larger deconstruction sequence). The tool can also be used for joint construction, where the instructor or learners can copy and paste their work into the Engagement Analyzer and progressively change parts of it so that it has similar engagement strategies as they see in the model. Finally, the instant visualization feature will help individual construction and the feedback provision as it can annotate the user-input text without relying on teacher's time and resources. The tool is thus compatible with a range of pedagogical activities in teaching writing (see Lesson ideas compatible with the notion of engagement, Brisk, 2020, Chapter 4). One important caveat is that the present version of the tool may not produce as accurate annotation as desired for self-directed learning scenes. However, it is hoped that more research—reliable annotated data and more appropriate machine learning architectures—will produce more accurate version of the tool that can be used by learners without supervision of experts, widening the potential pedagogical application of the tool.

6.3.5 Automated Writing Evaluation

As mentioned in Chapter 1, most of the widely used Automated Writing Evaluation engines focus extensively on lexical and grammatical aspects, neglecting discourse features (Attali, 2007; Carr, 2013; Enright & Quinlan, 2010; Lu, 2021). Thus, computational linguists and assessment researchers have called for automated analysis tools to analyze the discourse features of L2 language use (see Lu, 2021). This was one of the motivations for developing the end-to-end NLP system in this study. As demonstrated in Chapter 5, the features of engagement strategies and their diversity predicted writing scores alongside existing linguistic features at the levels of lexis (e.g., Kyle et al., 2018), syntax (Kyle & Crossley, 2017, 2018), and cohesion (Crossley et al., 2016, 2019). This finding is important because including engagement features

may improve the predictive accuracy of trained human raters. This also has important implications for the construct coverage of AWE systems. In fact, adding rhetorical features have been a much-needed update to AWE systems (Burstein et al., 2016; Carr, 2013; Lu, 2021), and the features of engagement strategies and diversity may fulfill such needs. Future research may examine how the combinations of these features impact on the assessment use argument of AWE engines, particularly in terms of enhancing Explanation and Extrapolation inferences (Chapelle et al., 2008; Enright & Quinlan, 2010), and the meaningfulness and generalizability of scores (Bachman & Palmer, 2010).

6.4 Limitations

There are a number of limitations to this dissertation study that need to be addressed in future studies. From Study 1 (see Chapter 3), two important limitations are worth reiterating. First, although the Engagement Discourse Treebank is the largest dataset of human-coded engagement resource analysis, the corpus size is still relatively small (i.e., 126,000 words; 4,600 sentences). For this reason, a few engagement strategies—CONCUR, ENDORSE, and PRONOUNCE—were extremely rare, and the current version of EDT possibly miss some infrequent patterns of lexico-grammatical realizations of these types. Although the impact of missing lexico-grammatical patterns is of slightly less concern because of the use of large pre-trained language model (i.e., RoBERTa, Liu et al., 2019), which may help generalize the pattern from scarce realization, there is still concerns that training data may not have enough coverage of different kinds of "seeds" for adequate generalization. Therefore, increasing the corpus size is a top priority to enhance the performance of the Engagement Analyzer. Second, the minimal contextual approach used in corpus sampling may limit the scope of annotation. Since neither the

annotators nor the machine learning system looked at contexts beyond three-sentence segments, their decisions on this category were based on reduced contextual information. In a future study, annotations based on paragraphs may further enhance the validity and reliability of annotation and machine learning systems. At the same time, this limitation resulted from a necessary compromise. A dataset based on paragraphs as the unit of analysis would have significantly reduced the number of writing samples (as a proxy for writing styles, topics, genres, discourse stages in writing, etc.) represented in the annotation dataset, and this was avoided to maximize the effectiveness of the random sampling procedure. The reduction of contexts was not necessarily a critical issue, given the current state-of-the-art natural language processing techniques, because the Transformer models used in the current study limited the window size to 512 sub-tokens (approximately 384 words depending on morphological complexity). Overall, given the promising results of the current study, future studies should aim to sample larger units of writing samples—paragraphs—and provide more contextually aware annotations and test whether the proposed Engagement Analysis pipelines can still approximate to human annotation.

Relatedly, the moderate inter-coder agreement in the annotation reinforces the view that the original engagement framework and their category explanation may be insufficient to arrive at highly reliable and replicable annotation (Fuoli, 2018). As fully elaborated in limitation section of Chapter 3, a high-quality annotation may require helps of experts in discourse analysis and/or domain experts who knows the disciplinary conventions. The development of workflow to achieve transparent and highly reliable annotation (e.g., Fuoli, 2018) merits further research in the domain of applied NLP research.

Study 2 investigated the distribution of engagement strategies across registers of university writing. One important limitation should be reiterated here (see other limitations in

Chapter 4). While Study 2 demonstrated the overall impacts of genre family, discipline, writer, level, and grades via a Bayesian MANOVA, the sheer number of levels of these factors hindered detailed a post hoc comparison as a part of the main analysis. Although a visual representation of the post hoc comparison was presented in the online supplementary material (https://osf.io/dvyem/?view_only=7854a6e80f804740a3beac6fd36f6a17), more in-depth descriptions by genre family and/or discipline will contribute to a more precise understanding of register patterns in written university assignments. Another important limitation includes the fact that no single NLP pipelines were perfect in identifying engagement strategies. To this end, I employed four different version of the pipeline and statistically controlled for their variability in order to avoid overconfidence in the reported results. The study also found that the effects of selected models were negligible according to the inter-item correlation coefficients (> .9) as well as multilevel MANOVA analysis. Thus, the selection of particular models may be negligible in the current study. It can be concluded that the top priority remains increasing the size of the annotation corpus and enhancing the quality of it.

Study 3 demonstrated the contribution of automatically identified engagement features to a predictive model for writing scores on a standardized English proficiency exam at the levels of lexis, syntax, and cohesion. Although the study demonstrated the utility of evenness indices in describing the diversity of closed-set Engagement strategies, the stability of this newly adapted index still needs to be tested regarding its intrinsic relation to text length. As research on lexical diversity indices has enjoyed careful analyses of the impact of text length on existing measures (Koizumi & In'nami, 2012; Zenker & Kyle, 2021), the same type of analysis may allow researchers to establish the text-length stability of engagement diversity measures. Another important limitation of Study 3 was that I only examined reduced sets of lexical, syntactic, and

cohesion indices and tested their relative predictive validity through observed variable regression analysis. Although this approach does not undermine the validity of the findings, it may be beneficial to conduct latent variable regression analysis using Structural Equation Modeling (SEM) to see how each latent construct of lexis, syntax, cohesion, and the rhetorical features of engagement together explain the variance in writing scores. Analyses based on SEM also allow the formal evaluation of whether there is measurement invariance of relative contributions of lexical, syntactic, cohesion, and rhetorical features across task types. Finally, the same limitation and possible backing against this limitation regarding the selection of particular versions of Engagement analyzer is relevant to Study 3.

6.5 Conclusion

The overarching goals of the current dissertation project were twofold. First, it aimed to create an automated system to conduct linguistic annotation of rhetorical discourse features that have predominantly been investigated through in-depth qualitative discourse analysis (Lancaster, 2014; Martin & White, 2005; Ryshina-Pankova, 2014; Wu, 2007; Xu & Nesi, 2019). Second, it aimed to showcase the use cases of the automated system developed in two contexts—register analysis of university written assignments and task-based writing assessment in a standardized English proficiency exam setting. To this end, Chapter 3 (Study 1) of this study introduced the development of a human-annotated corpus of academic written English, which draws on a system of engagement in the Appraisal framework (Martin & White, 2005). The outcome of this effort, the Engagement Discourse Treebank (EDT), was then used to train an end-to-end machine learning system to identify spans and discourse categories of engagement enacting resources. The best ML model, using RoBERTa embedding + LSTM layers, performed nearly as well as (or

even outperformed) the baseline of trained human annotators. Study 1 concluded with some recommended procedures for similar studies to develop end-to-end machine learning models for tailored linguistic analysis. A series of machine learning models developed in Study 1 were used to investigate the register of university written assignments in Study 2 (Chapter 4) and writing performance in the context of a standardized L2 English exam in Study 3 (Chapter 5). Both studies highlighted the benefits of incorporating automated engagement analysis in the respective goals of the research. For example, the findings of the register analysis (Study 2) highlighted the interplay of genre family and disciplines in determining the relative occurrences of different engagement strategies. The analysis of writing exam responses (Study 3) revealed that rhetorical features operationalized as diversity in engagement strategies can complement a predictive model of assessed writing scores, along with existing linguistic measures at the levels of lexis, syntax, and cohesion. Given the promising findings obtained the three studies, I conclude the present dissertation study with a call for larger-scale collaborations between discourse analysts, SLA researchers, assessment specialists, and computational linguists to develop automated linguistic annotation systems that can undertake large-scale analyses of less commonly investigated but important discourse features for research and education purposes (Lu, 2021). Such collaborative endeavor will result in useful methodological frameworks to innovate the research and pedagogy of Instructed Second Language Acquisition (Spada, 2021), English for Academic Purposes (Xie, 2020), writing instruction (Ferris & Hedgcock, 2013; Sparks et al., 2014), and next-generation Automated Writing Evaluation (Burstein et al., 2016).

REFERENCES

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. John Benjamins Pub. Co.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, Inc.

Aguiar, G., Krawczyk, B., & Cano, A. (2022). *A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework* (arXiv:2204.03719). arXiv. http://arxiv.org/abs/2204.03719

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, *4*(1), 71–83. https://doi.org/10.3366/E1749503209000227

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*, *2007*(1), i–22. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Bachman, L. F., & Palmer, A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. *TESOL Quarterly*, *16*(4), 449. https://doi.org/10.2307/3586464

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bakhtin, M. M. (Mikhail M. (1981). *The dialogic imagination: Four essays* (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.). University of Texas Press.

Bax, S., Nakatsuhara, F., & Waller, D. (2019). Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels. *System*, *83*, 79–95. https://doi.org/10.1016/j.system.2019.02.010

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *ArXiv Preprint ArXiv:2004.05150*.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016). Universal Dependencies for Learner English. *ArXiv:1605.04278 [Cs]*. http://arxiv.org/abs/1605.04278

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. https://doi.org/10.1016/j.jslw.2014.09.004

Biber, D. (1984). *A Model of Textual Relations within the Written and Spoken Modes* [Unpublished doctoral dissertation]. University of Southern California.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers*. J. Benjamins.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. *ETS TOEFL Monograph Series*, *25*.

Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, *11*(1), 1–34. https://doi.org/10.1080/01638538809544689

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse*, *9*(1). https://doi.org/10.1515/text.1.1989.9.1.93

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. http://www.jstor.org/stable/41307614

Biber, D., Gray, B., & Staples, S. (2014). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, 100869. https://doi.org/10.1016/j.jeap.2020.100869

Biber, D., Johansso, S., Leech, G., Conrad, S., & Finegan, E. (Eds.). (1999). *Longman grammar of spoken and written English* (10. impression). Longman.

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of Spoken and Written English*. John Benjamins Publishing Company. https://doi.org/10.1075/z.232

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A Corpus of Non-Native English. *ETS Research Report Series*, *2013*(2), i–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns*. Routledge.

Bouziri, B. (2021). A tripartite interpersonal model for investigating metadiscourse in academic lectures. *Applied Linguistics*, *42*(5), 970–989. https://doi.org/10.1093/applin/amab001

Brisk, M. E. (2020). *Language in writing instruction: Enhancing literacy in grades 3-8*. Routledge.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Burstein, J., Elliot, N., Klebanov, B. B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing MentorTM: Writing Progress Using Self-Regulated Writing Support. *The Journal of Writing Analytics*, *2*(1), 285–313. https://doi.org/10.37514/JWA-J.2018.2.1.12

Burstein, J., Elliot, N., & Molloy, H. (2016). Informing Automated Writing Evaluation Using the Lens of Genre: Two Studies. *CALICO Journal*, *33*(1), 117–141. https://doi.org/10.1558/cj.v33i1.26374

Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, *1*, 1–47. https://doi.org/10.1093/applin/1.1.1

Carr, N. T. (2013). Computer-automated scoring of written responses. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1063–1078). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla124

Chafe, W. L., & Nichols, J. (Eds.). (1986). *Evidentiality: The linguistic coding of epistemology*. Ablex Pub. Corp.

Chang, P., & Schleppegrell, M. (2011). Taking an effective authorial stance in academic writing: Making the linguistic resources explicit for L2 writers in the social sciences. *Journal of English for Academic Purposes*, *10*(3), 140–151. https://doi.org/10.1016/j.jeap.2011.05.005

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. SAGE Publications, Inc.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Routledge. https://doi.org/10.4324/9780203937891

Charles, M. (2006). The Construction of Stance in Reporting Clauses: A Cross-disciplinary Study of Theses. *Applied Linguistics*, *27*(3), 492–518. https://doi.org/10.1093/applin/aml021

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. *ArXiv:1906.04341 [Cs]*. http://arxiv.org/abs/1906.04341

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cotos, E. (2014). *Genre-Based Automated Writing Evaluation for L2 Research Writing*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137333377

Cotos, E., & Pendar, N. (2015). Discourse classification into rhetorical functions for AWE feedback. *CALICO Journal*, *0*(0). https://doi.org/10.1558/cj.v33i1.27047

Council of Europe (Ed.). (2020). *Common European framework of reference for languages: Learning, teaching, assessment ; companion volume*. Council of Europe Publishing.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *ArXiv Preprint ArXiv:1901.02860*.

Daly, A., Baetens, J., & De Baets, B. (2018). Ecological Diversity: Measuring the Unmeasurable. *Mathematics*, *6*(7), 119. https://doi.org/10.3390/math6070119

De Oliveira, L. C., & Schleppegrell, M. (2015). *Focus on grammar and meaning*. Oxford University Press.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Derewianka, B. (2007). Using Appraisal Theory to Track Interpersonal Development in Adolescent Academic Writing. In *Advances in Language and Education*. Bloomsbury Academic. https://doi.org/10.5040/9781474212045

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dror, R., Peled-Cohen, L., Shlomov, S., & Reichart, R. (2020). *Statistical Significance Testing for Natural Language Processing*. Morgan & Claypool.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. https://www.aclweb.org/anthology/W16-4011

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). 10. Assessing writing. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 147–164). De Gruyter. https://doi.org/10.1515/9781614513827-012

Eggins, S. (2004). *An introduction to systemic functional linguistics* (2nd ed). Continuum.

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. https://doi.org/10.1111/modl.12637

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford University Press.

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. https://doi.org/10.1177/0265532210363144

Ferris, D. R., & Hedgcock, J. (2013). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). Routledge.

Fiacco, J., Jiang, S., Adamson, D., & Rosé, C. (2022). Toward Automatic Discourse Parsing of Student Writing Motivated by Neural Interpretation. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 204–215. https://doi.org/10.18653/v1/2022.bea-1.25

Fuoli, M. (2018). A stepwise method for annotating appraisal. *Functions of Language*, *25*(2), 229–258. https://doi.org/10.1075/fol.15016.fuo

Fuoli, M., & Hommerberg, C. (2015). Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, *10*(3), 315–349. https://doi.org/10.3366/cor.2015.0080

Gabry, J., & Češnovar, R. (2021). cmdstanr: R Interface to'CmdStan'. *URL: Https://Mc-Stan. Org/Cmdstanr, Https://Discourse. Mc-Stan. Org*.

Gelman, A. (2005). Analysis of variance—Why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53. https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-1/Analysis-of-variancewhy-it-is-more-important-than-ever/10.1214/009053604000001048.full

Gelman, A. (Ed.). (2014). *Bayesian data analysis* (2nd ed). Chapman & Hall/CRC.

Gelman, A. (2016). Prior Choice Recommendations. *Stan Github Page*. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. *ArXiv:2011.01808 [Stat]*. http://arxiv.org/abs/2011.01808

Gilabert, R., & Barón, J. (2018). Chapter 7. Independently measuring cognitive complexity in task design for interlanguage pragmatics development. In N. Taguchi & Y. Kim (Eds.), *Task-Based Language Teaching* (Vol. 10, pp. 160–190). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.10.07gil

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). *Maxout Networks* (arXiv:1302.4389). arXiv. http://arxiv.org/abs/1302.4389

Goulart, L., & Wood, M. (2021). Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science, 6*, 107–137.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Gray, B., & Biber, D. (2012). Current conceptions of stance. In *Stance and voice in written academic genres* (pp. 15–33). Springer.

Gray, B., Geluso, J., & Nguyen, P. (2019). The Longitudinal Development of Grammatical Complexity at the Phrasal and Clausal Levels in Spoken and Written Responses to the *TOEFL iBT*® Test. *ETS Research Report Series*, *2019*(1), 1–51. https://doi.org/10.1002/ets2.12280

Gu, W., Zheng, B., Chen, Y., Chen, T., & Van Durme, B. (2022). *An Empirical Study on Finding Spans* (arXiv:2210.06824). arXiv. http://arxiv.org/abs/2210.06824

Gudmestad, A., House, L., & Geeslin, K. L. (2013). What a Bayesian Analysis Can Do for SLA: New Tools for the Sociolinguistic Study of Subject Expression in L2 Spanish: Bayesian Analysis for SLA. *Language Learning*, *63*(3), 371–399. https://doi.org/10.1111/lang.12006

Hagiwara, M. (2022). *Real-world natural language processing*. Manning Publications.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *An introduction to functional grammar* (4th ed). Routledge.

Heck, R. H., & Thomas, S. L. (2020). *An introduction to multilevel modeling techniques: MLM and SEM approaches* (Fourth edition). Routledge.

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Honnibal, M., Ines, M., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* (3.3). https://spacy.io

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* [Python]. https://doi.org/10.5281/zenodo.1212303 (Original work published 2014)

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. Palgrave Macmillan.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Hox, J. J. (2018). *Multilevel Analysis: Techniques and Applications* (Third edition). Routledge.

Hunston, S. (2000). Evaluation and the planes of discourse: Status and value in persuasive texts. In S. Hunston & G. Thompson (Eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse* (pp. 176–207). Oxford University Press, UK.

Hunston, S. (2004). Counting the uncountable: Problems of identifying evaluation in a text and in a corpus. In A. S. Partington (Ed.), *Corpora and Discourse*.

Hunston, S., & Thompson, G. (2000). *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press, UK.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. *NCTE Research Report No. 3*. http://eric.ed.gov/?id=ED113735

Hyland, K. (2005a). *Metadiscourse: Exploring interaction in writing*. Continuum.

Hyland, K. (2005b). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173–192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics*, *113*, 16–29. https://doi.org/10.1016/j.pragma.2017.03.007

Hyland, K., & Jiang, F. (Kevin). (2022). Metadiscourse choices in EAP: An intra-journal study of JEAP. *Journal of English for Academic Purposes*, *60*, 101165. https://doi.org/10.1016/j.jeap.2022.101165

Hymes, D. (1972). On Communicative Competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics.* (pp. 269–293). Penguin.

Ishikawa, S. (2013). *The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian Learners of English*. 神戸大学国際コミュニケーションセンター. https://doi.org/10.24546/81006678

Ishikawa, S. (2018). The ICNALE Edited Essays; A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Studies*, *25*, 117–130.

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning*, *63*(S1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–43). John Benjamins Publishing Company.

Jiang, J., & Zhai, C. (2007). Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 264–271. https://aclanthology.org/P07-1034

Jiang, Z., Xu, W., Araki, J., & Neubig, G. (2020). Generalizing Natural Language Analysis through Span-relation Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2120–2133. https://doi.org/10.18653/v1/2020.acl-main.192

Kellogg, R. T. (1996). A model of working memory in writing. In *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–71). Lawrence Erlbaum Associates, Inc.

Kershaw, D., & Koeling, R. (2020). *Elsevier OA CC-BY Corpus* [Data set]. Mendeley. https://doi.org/10.17632/ZM33CDNDXS.3

Kim, M. M., Crossley, S. A., & 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. https://doi.org/10.1111/modl.12447

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564.

Krebs, C. J. (1999a). *Ecological Methodology*. Benjamin/Cummings.

Krebs, C. J. (1999b). Species Diversity Measures. In *Ecological methodology* (pp. 598-).

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, JAGS, and stan*. Academic Press.

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*.

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing*, *45*, 100467. https://doi.org/10.1016/j.asw.2020.100467

Kyle, K., Choe, A. T., Eguchi, M., LaFlair, G., & Ziegler, N. (2021). A Comparison of Spoken and Written Language Use in Traditional and Technology-Mediated Learning Environments. *ETS Research Report Series*, *2021*(1), 1–29. https://doi.org/10.1002/ets2.12329

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *Modern Language Journal*. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 1–17. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using words, n-gram, and dependency bigram indices. In S. Granger (Ed.), *Perspectives on the Second Language Phrasicon: The View from Learner Corpora.* Multiligual Matters.

Kyle, K., Eguchi, M., Choe, A. T., & LaFlair, G. (2022). Register variation in spoken and written language use across technology-mediated and non-technology-mediated learning environments. *Language Testing*, *39*(4), 618–648. https://doi.org/10.1177/02655322211057868

Kyle, K., & Sung, H. (2023). An Argument Structure Construction Treebank. *The First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, 51–62.

Labov, W. (1984). Intensity. In D. Schiffrin (Ed.), *Meaning, form and use in context: Linguistic applications* (pp. 43–70). Georgetown Univ. Pr.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference: A case study of a high-stakes speaking assessment. *Language Testing*, *34*(4), 451–475. https://doi.org/10.1177/0265532217713951

Lam, S. L., & Crosthwaite, P. (2018). APPRAISAL resources in L1 and L2 argumentative essays: A contrastive learner corpus-informed study of evaluative stance. *Journal of Corpora and Discourse Studies*, *1*(1), 8. https://doi.org/10.18573/jcads.1

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE.

Lancaster, Z. (2014). Exploring valued patterns of stance in upper-level student writing in the disciplines. *Written Communication*, *31*(1), 27–57. https://doi.org/10.1177/0741088313515170

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

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

Laughlin, V. T., Wain, J., & Schmidgall, J. (2015). Defining and operationalizing the construct of pragmatic competence: Review and recommendations. *ETS Research Report Series*, *2015*(1), 1–43. https://doi.org/10.1002/ets2.12053

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 188–197. https://doi.org/10.18653/v1/D17-1018

Lee, S. H. (2017). Use of implicit intertextuality by undergraduate students: Focusing on Monogloss in argumentative essays. *Linguistics & the Human Sciences*, *13*(1), 150–178. https://doi.org/10.1558/lhs.30651

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. *ArXiv:1907.11692 [Cs]*. http://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. https://doi.org/10.1075/ijcl.15.4.02lu

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. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2021). Directions for future automated analyses of L2 written texts. In R. M. Manchón & C. Polio, *The Routledge handbook of second language acquisition and writing* (1st ed., pp. 370–382). Routledge. https://doi.org/10.4324/9780429199691-36

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60).

Mackey, A., & Gass, S. M. (2018). *Second language research: Methodology and design* (2nd ed.). Routledge.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Martin, J. R., Matthiessen, C. M. I. M., & Painter, C. (1997). *Working with functional grammar*. Arnold.

Martin, J. R., & Rose, D. (2007). *Working with discourse: Meaning beyond the clause* (2nd ed.). Continuum.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Mauranen, A., & Bondi, M. (2003). Evaluative language use in academic discourse. *Journal of English for Academic Purposes*, *2*(4), 269–271.

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. https://doi.org/10.1177/00238309010440030101

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor and Francis, CRC Press.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Handbook of Instructed Second Language Acquisition* (pp. 50–68). Routledge.

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency. *Instructed Second Language Acquisition*, *3*(2). https://doi.org/10.1558/isla.38248

Miller, G. (1995). *Wordnet*. *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Misra, D. (2020). *Mish: A Self Regularized Non-Monotonic Activation Function* (arXiv:1908.08681). arXiv. http://arxiv.org/abs/1908.08681

Mizumoto, A. (2022). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, lang.12518. https://doi.org/10.1111/lang.12518

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S., & Rillig, M. C. (2014). Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*, *4*(18), 3514–3524. https://doi.org/10.1002/ece3.1155

Nalborczyk, L., Batailler, C., Vilain, A., & Bürkner, P.-C. (2018). *An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian*. 18.

Nesi, H. (2021). Sources for courses: Metadiscourse and the role of citation in student writing. *Lingua*, *253*, 103040. https://doi.org/10.1016/j.lingua.2021.103040

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press. https://doi.org/10.1017/9781009030199

Nesi, H., & Gardner, S. (2018). The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing*, *38*, 51–55. https://doi.org/10.1016/j.asw.2018.06.005

Nini, A. (2019). The multi-dimensional analysis tagger. *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67–94.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian Revolution in Second Language Research: An Applied Approach: Bayesian Revolution in L2 Research. *Language Learning*, *68*(4), 1032–1075. https://doi.org/10.1111/lang.12310

Norouzian, R., Miranda, M. D., & Plonsky, L. (2019). A Bayesian Approach to Measuring Evidence in L2 Research: An Empirical Investigation. *The Modern Language Journal*, *103*(1), 248–261. https://doi.org/10.1111/modl.12543

Norris, J. M. (2015). Statistical Significance Testing in Second Language Research: Basic Problems and Suggestions for Reform: Statistical Significance Testing in Second Language Research. *Language Learning*, *65*(S1), 97–126. https://doi.org/10.1111/lang.12114

Norris, J. M., & Ortega, L. (2003). Defining and Measuring SLA. In C. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition*.

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. https://doi.org/10.1093/applin/amp044

Ochs, E., & Schieffelin, B. (1989). Language has a heart. *Text - Interdisciplinary Journal for the Study of Discourse*, *9*(1). https://doi.org/10.1515/text.1.1989.9.1.7

Palmer, F. R. (2001). *Mood and modality* (2nd ed). Cambridge University Press.

Papay, S., Klinger, R., & Padó, S. (2020). Dissecting Span Identification Tasks with Performance Prediction. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4881–4895. https://doi.org/10.18653/v1/2020.emnlp-main.396

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research: Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Polio, C., & Yoon, H.-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*(1), 165–188. https://doi.org/10.1111/ijal.12200

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *ArXiv Preprint ArXiv:1806.03822*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *ArXiv Preprint ArXiv:1606.05250*.

Ramponi, A., & Plank, B. (2020). *Neural Unsupervised Domain Adaptation in NLP---A Survey* (arXiv:2006.00632). arXiv. https://doi.org/10.48550/arXiv.2006.00632

Rao, A. R. (2022). ASRtrans at SemEval-2022 Task 4: Ensemble of Tuned Transformer-based Models for PCL Detection. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 344–351. https://doi.org/10.18653/v1/2022.semeval-1.44

Read, J., & Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, *46*(3), 421–447. https://doi.org/10.1007/s10579-010-9135-7

Revesz, A., Ekiert, M., & Torgersen, E. N. (2014). The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance. *Applied Linguistics*. https://doi.org/10.1093/applin/amu069

Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*(2), 159–177.

Römer, U., & Swales, J. M. (2010). The Michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes*, *9*(3), 249.

Rose, D., & Martin, J. R. (2012). *Learning to write, reading to learn: Genre, knowledge, and pedagogy in the Sydney School*. Equinox Pub.

Ryshina-Pankova, M. (2014). Exploring academic argumentation in course-related blogs through ENGAGEMENT. In G. Thompson & L. Alba-Juez (Eds.), *Pragmatics & Beyond New Series* (Vol. 242, pp. 281–302). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.242.14rys

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*, 103–126. https://doi.org/10.1037/met0000275

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). Workflow Techniques for the Robust Use of Bayes Factors. *ArXiv:2103.08744 [Stat]*. http://arxiv.org/abs/2103.08744

Schneer, D. (2014). Rethinking the Argumentative Essay. *TESOL Journal*, *5*(4), 619–653. https://doi.org/10.1002/tesj.123

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Silverman, D. (2018). *Qualitative Research* (5th ed.).

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). *Stanford Sentiment Treebank*. Stanford University. https://nlp. stanford. edu/sentiment/treebank.

Spada, N. (2021). Reflecting on task-based language teaching from an Instructed SLA perspective. *Language Teaching*, 1–13. https://doi.org/10.1017/S0261444821000161

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment. *ETS Research Report Series*, *2014*(2), 1–52. https://doi.org/10.1002/ets2.12035

Staples, S., Biber, D., & Reppen, R. (2018). Using Corpus-Based Register Analysis to Explore the Authenticity of High-Stakes Language Exams: A Register Comparison of TOEFL iBT and Disciplinary Writing Tasks. *The Modern Language Journal*, *102*(2), 310–332. https://doi.org/10.1111/modl.12465

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Swales, J. M., & Feak, C. B. (2000). *English in today's research world: A writing guide*. University of Michigan Press.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills* (3rd ed.). University of Michigan Press Ann Arbor, MI.

Taguchi, N., Fernández, L., & Jiang, Y. (2021). 2 Systemic functional linguistics applied to analyze L2 speech acts: Analysis of advice-giving in a written text. In J. C. Félix-Brasdefer & R. Shively (Eds.), *New Directions in Second Language Pragmatics* (pp. 27–57). De Gruyter. https://doi.org/10.1515/9783110721775-006

Taguchi, N., & Kim, Y. (Eds.). (2018). *Task-Based Approaches to Teaching and Assessing Pragmatics* (Vol. 10). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.10

Thampi, A. (2022). *Interpretable AI*. Manning Publications.

Thompson, G., & Alba-Juez, L. (Eds.). (2014). *Evaluation in context*. John Benjamins Publishing Company.

Thompson, G., & Hunston, S. (2000). Evaluation: An introduction. In S. Hunston & G. Thompson (Eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse* (pp. 1–27). Oxford University Press, UK.

Tunstall, L. (2022). *Natural Language Processing with Transformers*. 417.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. https://mc-stan.org/loo/

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*. https://doi.org/10.1214/20-ba1221

Wang, X., & Wang, Y. (2022). Sentence-Level Resampling for Named Entity Recognition. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2151–2165. https://doi.org/10.18653/v1/2022.naacl-main.156

White, P. R. R. (2003). Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text - Interdisciplinary Journal for the Study of Discourse*, *23*(2). https://doi.org/10.1515/text.2003.011

Williams, A., Nangia, N., & Bowman, S. R. (2018). *The multi-genre nli corpus*.

Winter, B., & Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, *15*(11), e12439. https://doi.org/10.1111/lnc3.12439

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. http://arxiv.org/abs/1910.03771

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

Wu, S. M. (2007). The use of engagement resources in high- and low-rated undergraduate geography essays. *Journal of English for Academic Purposes*, *6*(3), 254–271. https://doi.org/10.1016/j.jeap.2007.09.006

Xie, J. (2020). A review of research on authorial evaluation in English academic writing: A methodological perspective. *Journal of English for Academic Purposes*, *47*, 100892. https://doi.org/10.1016/j.jeap.2020.100892

Xu, X., & Nesi, H. (2017). An analysis of the evaluation contexts in academic discourse. *Functional Linguistics*, *4*(1). https://doi.org/10.1186/s40554-016-0037-x

Xu, X., & Nesi, H. (2019). Differences in engagement: A comparison of the strategies used by British and Chinese research article writers. *Journal of English for Academic Purposes*, *38*, 121–134. https://doi.org/10.1016/j.jeap.2019.02.003

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189. https://aclanthology.org/P11-1019

Yasuda, S. (2015). Exploring changes in FL writers' meaning-making choices in summary writing: A systemic functional approach. *Journal of Second Language Writing*, *27*, 105–121. https://doi.org/10.1016/j.jslw.2014.09.008

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. https://www.aclweb.org/anthology/P13-4001

Yoon, H.-J. (2017a). Textual voice elements and voice strength in EFL argumentative writing. *Assessing Writing*, *32*, 72–84. https://doi.org/10.1016/j.asw.2017.02.002

Yoon, H.-J. (2017b). Textual voice elements and voice strength in EFL argumentative writing. *Assessing Writing*, *32*, 72–84. https://doi.org/10.1016/j.asw.2017.02.002

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., & Yang, L. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhu, Q., Lin, Z., Zhang, Y., Sun, J., Li, X., Lin, Q., Dang, Y., & Xu, R. (2021). HITSZ-HLT at SemEval-2021 Task 5: Ensemble Sequence Labeling and Span Boundary Detection for Toxic Span Detection. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 521–526. https://doi.org/10.18653/v1/2021.semeval-1.63