EXPLORING USES OF AUTOMATED ESSAY SCORING FOR ESL

Geneva Marie Tesh

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

July 2023

APPROVED:

Youngjin Lee, Major Professor
S. Willard Elieson, Committee Member
Regina Kaplan-Rakowski, Committee Member
Yunjo An, Chair of the Department of Learning Technologies
Kinshuk, Dean of the College of Information
Victor Prybutok, Dean of the Toulouse Graduate School

Tesh, Geneva Marie. *Exploring Uses of Automated Essay Scoring for ESL*. Doctor of Philosophy (Learning Technologies), July 2023, 146 pp., 14 tables, 28 figures, 2 appendices, references, 104 titles.

Manually grading essays and providing comprehensive feedback pose significant challenges for writing instructors, requiring subjective assessments of various writing elements. Automated essay scoring (AES) systems have emerged as a potential solution, offering improved grading consistency and time efficiency, along with insightful analytics. However, the use of AES in English as a Second Language (ESL) remains rare. This dissertation aims to explore the implementation of AES in ESL education to enhance teaching and learning.

The dissertation presents a study involving ESL teachers who learned to use a specific AES system called LightSide, a free and open text mining tool, to enhance writing instruction. The study involved observations, interviews, and a workshop where teachers learned to build their own AES using LightSide. The study aimed to address questions related to teacher interest in using AES, challenges faced by teachers, and the influence of the workshop on teachers' perceptions of AES. By exploring the use of AES in ESL education, this research provides valuable insights to inform the integration of technology and enhance the teaching and learning of writing skills for English language learners.

ACKNOWLEDGMENTS

I am deeply grateful to the network that supported me throughout the completion of this dissertation. Without my family's unending support and encouragement, I would not have been able to achieve this milestone in my academic journey. My parents instilled in me a passion for learning, and my children, Mazzen, Ramzy, Isabel, and Zachary, have been my greatest inspiration. My husband, Scott, has been my constant cheerleader, taking excellent care of our family, and being my best friend and life partner.

My heartfelt appreciation goes to my dissertation chair, Dr. Youngjin Lee, and committee members Dr. Bill Elieson and Dr. Regina Kaplan-Rakowski. Dr. Lee's patient guidance and suggestions were invaluable in helping me navigate the dissertation process. Dr. Elieson's mentorship and weekly chats kept me on track from the very first day I began this program. Dr. Kaplan-Rakowski provided much-needed feedback throughout the journey.

I'm deeply honored and privileged to have been part of the LTEC doctoral program at the University of North Texas, learning from leading experts in the Learning Technologies field. I feel fortunate to have been part of an exceptional cohort of scholars, and I am grateful for their unwavering support, encouragement, and motivation throughout the trials and tribulations of earning a Ph.D. I'm especially grateful to my dear friends Kristi Larson and Roger Chambers, who generously checked my data and helped me cross the finish line.

Finally, I would like to express my gratitude to all the ESL teachers who participated in this study. Their assistance and insights were invaluable, and without their cooperation, this research would not have been possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AES	Automated essay scoring
AEG	Automated writing evaluation
AI	Artificial intelligence
ALPAC	Automatic language processing advisory committee
ASAP	Automated student assessment prize
AWE	Automated writing evaluation
CAE	Certificate in advanced English
CALL	Computer-assisted language learning
CET	College English test
CLC	Cambridge Learner Corpus
EFL	English as a foreign language
ELL	English language learner
ESL	English as a second language
ENA	Epistemic network analysis
ETS	Educational testing services
FCE	First certificate in English
ICLE	International corpus of learner English
IEA	Intelligent essay assessor
IEP	Intensive English program
LA	Learning analytics
LSA	Latent semantic analysis

ML	Machine learning
MT	Machine translation
NLP	Natural language processing
NNS	Non-native speakers of English
PEG	Project essay grade
TESOL	Teachers of English to speakers of other languages
TOEFL	Test of English as a foreign language
TWE	Test of written English

CHAPTER 1

INTRODUCTION

Manually grading essays and providing quality feedback is arguably one of the most challenging tasks for writing instructors. Judging writing quality is by nature a subjective process that involves not merely an assessment of correctness but also careful consideration of style, flow, wit, creativity, critical thinking, and linguistic sophistication. Beyond the issues of time constraints and human subjectivity, graders may provide inconsistent feedback due to grading fatigue and distractions. Yet frequent, corrective feedback is essential for students striving to improve their writing skills. Feedback is especially important to English language learners, who are not only learning the standard conventions of writing but are also in the process of acquiring vocabulary and mastering grammar rules. With the potential to improve grading consistency and reduce the amount of time needed to score and deliver feedback to students, automated essay scoring (AES) systems may offer a solution. AES systems also provide helpful analytics that identify recurring patterns and errors in writing samples, affording teachers special insight in analyzing their students' writing and offering much more targeted feedback and focused writing instruction. Despite these affordances, the use of AES in English as a Second Language (ESL) and English as a Foreign Language (EFL) remains rare.

AES systems have undergone extensive research in validity, reliability, efficiency, usage, and perception, with very positive results. The correlation between a human score and a machine score is as strong as the correlation between two human scores (Klebanov & Madnani, 2022). However, most published studies on the topic have centered around education and testing environments in the United States, under the assumption that the writers are native or

native-like speakers of English (Hyland & Hyland, 2019). In terms of learning and communicating in English, non-native speakers (NNSs) of English outnumber native speakers, as English language proficiency has become a global endeavor and is considered a basic skill in many countries (Fleckenstein et al., 2016). Because English is often the language for global education and business, there are many contexts in which NNSs communicate in English. Among other reasons, some NNSs use English at work or school in their home countries, some work or study abroad, and some are immigrants or the children of immigrants in English-speaking countries. Given the ever-growing number of NNSs, that population possibly makes up the largest market for AES systems (Liang & Guo, 2020). Thus, the use of AES systems for non-native, or learner, English presents an important research opportunity.

Language teachers' perceptions towards integrated technology in general could hinder or prevent them from using AES. Through a review of 18 empirical studies on the effects of integrating technology in ESL/EFL writing instruction, Al-Wasy (2020) concluded that technology has an overall positive effect on the development of learners' writing proficiency. However, many English language teachers do not use technology in the classroom for a multitude of reasons, including inadequate access to tools, lack of knowledge or exposure, low confidence in using new tools, limited professional development and support, time or other practical constraints, and negative attitudes (Alamri, 2021). Such challenges may affect teachers' willingness to adopt AES in the classroom.

The purpose of this dissertation is to explore ways in which teachers can implement AES in ESL education to enhance teaching and learning. The first chapter provides background information about the development of AES systems and discusses the context of AES for English

learners. The second chapter synthesizes current research on AES systems in ESL education. The third and fourth chapters present a study on ESL teachers developing and using an AES system to enhance writing instruction. Study participants learned how to use Lightside, a free and open text mining tool, as an automated essay scoring system. The study consisted of observations and interviews with teachers who completed a brief workshop to learn how to use LightSide to build their own AES. The following questions guided the discussion:

- Are teachers interested in using AES to enhance teaching and learning?
- How easily can teachers learn to use LightSide?
- What challenges did teachers face in learning to use the platform?
- How do trained teachers intend to use LightSide to enhance teaching and learning?
- Did the LightSide workshop influence teachers' perceptions about AES?

The final chapter describes the implications of the study, its limitations, and the potential for future research.

Background

A Brief History of NLP

AES systems are made possible by natural language processing (NLP), the use of algorithms that allows machines to achieve human-like language processing for a variety of tasks or applications (Liddy, 2001). NLP research began in the 1940s with a machine translation (MT) project in 1946 attempting to break enemy codes during World War II (Liddy, 2001), and grew into an interdisciplinary field combining research from linguistics, computer science, and cognitive psychology. In the 1950s, NLP research continued to focus on MT, with the IBM-Georgetown Demonstration, a limited experiment of automatic translation from Russian to

English (Jones, 1994). During this time, research into speech recognition began, leading to the Bell Labs Audrey system for speech recognition (Klebanov & Madnani, 2022). Early work in NLP was largely empirical and data-driven (Lee, 2003). With the publication of *Syntactic Structures*, Noam Chomsky (1957) introduced the concept of generative grammar, providing new insight into how a computer could potentially understand human language. In 1958, John McCarthy developed the programming language LISP, which is still in use today, and in 1964, Weizenbaum created ELIZA, a typewritten comment and response program designed to replicate a patient's therapy session with a psychologist (Foote, 2019).

NLP researchers and developers in the 1950s and early 1960s were optimistic that machines would soon fully communicate with humans in natural language and be capable of producing automatic translations similar to what a human translator might produce, but this initial promise failed in 1966 with the establishment of ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science) and its 1966 report that MT was unachievable at that time (Liddy, 2001). The ALPAC report quelled much of the NLP research in the 1960s. Regardless, theoretical work in the field continued. Chomsky (1965) introduced the transformational model of linguistic competence, which received some backlash from linguists who argued that transformational grammar focused too heavily on syntax while overlooking semantics (Liddy, 2001). Some concepts and models developed in response to transformational generative grammar included Fillmore's (1968) case grammar, Quillian's (1968) semantic networks, Schank's (1972) conceptual dependency theory, Wilks's (1973) preference semantics, and Kay's (1979) functional grammar.

The enormous size and unrestrictive nature of natural language made it very difficult to

apply standard parsing approaches using symbolic, handwritten rules, and thus by the 1980s, NLP research circled back to its roots in data-driven, empirical methods, as researchers explored probabilistic language models using large, annotated collections of text (corpora) to train sophisticated machine-learning algorithms (Lee, 2003). By the 1990s, symbolic and statistical approaches proved to be complementary in solving many of the earlier problems and limitations in NLP, and the interaction of the two approaches, combined with advances in machine-learning, allowed computers to acquire linguistic information directly from corpora. This development is critical in understanding how most AES systems are built today. A large dataset of student writing samples allows an AES system to extract features and recognize syntactic patterns.

Syntax represents only one facet of language. A nonsensical sentence can follow all the syntactic rules but still be meaningless. Semantics represents the aspect of language concerned with meaning and logic. An important development in the late 1990s produced Latent Semantic Analysis (LSA), a technique in NLP that uses large corpora to imitate human language by analyzing the relationships between words and sentences (Landauer & Dumais, 1997). LSA works by aggregating all the contexts in which words appear in corpora and then constructing statistical analysis to determine word similarities and relationships (Klebanov & Madnani, 2022). A notable advancement in NLP, LSA is employed in a wide range of applications including internet searches, intelligent tutoring systems, and plagiarism detection software. Many AES systems today incorporate LSA technology (Vitartas et al., 2016).

Over the past two decades, advances in NLP have developed rapidly. Bengio et al. (2002) proposed a language model that uses neural networks, a machine learning technique that

attempts to copy the way living organisms learn from input stimuli. New deep learning models have since emerged, enabling machine learning to extract information from vast data sets and identify features for evaluating essays (Klebanov & Madnani, 2022). Prior to the emergence of deep learning models, the task of AES involved the use of handcrafted features combined with supervised learning, a process that demanded considerable human effort in crafting and implementing each feature (Tashu et al., 2022). The application of deep learning models to AES has allowed developers to transition from linear models that relied on handcrafted features to more advanced, nonlinear neural network models that rely on extensive data input (Tashu et al., 2022).

The fusion of NLP techniques with AI-based neural networks has enabled computers to capture more nuanced aspects of human language, including sentiment, intent, and discourse patterns (Keezhatta, 2019). This integration has facilitated the development of sophisticated algorithms that can analyze and process language in a manner that more closely resembles human understanding. As a result of these advancements, the original NLP goals of machine translation and speech recognition have been realized, as demonstrated by widely used applications such as Google Translate and virtual assistants such as Siri and Alexa, making NLP an integral part of our daily lives. Developments in NLP and AI have paved the way for innovative applications in diverse fields, such as sentiment analysis in social media monitoring, chatbots for customer support, and text summarization for efficient information consumption.

For language educators, NLP creates endless possibilities and exciting opportunities for enhancing teaching and learning. These advancements provide new insights into understanding how students acquire language. NLP can facilitate personalized learning experiences, as AI-

powered tools can analyze students' individual strengths and weaknesses, allowing educators to tailor their instruction accordingly. Moreover, NLP-driven applications can offer real-time feedback on students' written and spoken language, helping them to better identify and address areas for improvements. In addition, NLP can be used to develop language models that simulate natural conversations, offering students valuable practice in a safe and controlled environment. Overall, the fusion of NLP and AI technologies create tremendous potential for language education, enabling the development of innovative and effective teaching methods.

Educational Applications of NLP

Natural language processing can enhance learning technologies in many ways. The earliest educational application of NLP was automated essay scoring, with research appearing as early as the 1960s, not long after machine translation and speech recognition (Klebanov & Madnani, 2022). Ellis Page (1966) published a seminal paper discussing the need for and feasibility of an automated system for scoring student writing. AES followed the same path as other initial NLP applications, with great excitement over its potential in the 1960s, followed by a couple of decades of stagnation due to exorbitant costs and unavailable technology, before finally being implemented at a large-scale in the late 1990s when the technology finally caught up with the theory (Klebanov & Madnani, 2022).

While research in AES continues to progress rapidly, other ways to apply NLP to education present new research opportunities and offer novel educational affordances. Litman (2016) identified the following roles for NLP in education: enhancing language teaching and learning, building intelligent tutoring systems, and processing language from data sources such as MOOC forums or textbooks to deliver analytics that could support the needs of students and

teachers. The growing interest in utilizing NLP for educational purposes has led to the formation of numerous research communities, special interest groups, conferences, and symposiums dedicated to exploring innovative educational applications. Examples of such events and organizations include the annual Workshop on Innovative Use of NLP for Building Educational Applications, EDAppsNLP, the National Council on Measurement in Education Conference on NLP in Assessment, and NLP Education Conference. These forums facilitate the exchange of ideas, collaboration, and advancements in the field, furthering the potential impact of NLP on education.

NLP in Language Education

Although language teaching and learning were among the earliest pioneers in the application of NLP tools, their use in language classes has been relative limited. However, this situation is rapidly changing (Antoniadis & Desmet, 2019). Meurers (2019) identified the two general uses of NLP in language learning as the analysis of *learner* language, i.e., texts produced by students, and the analysis of *native* language, i.e., texts produced by native speakers of the target language. Analyzing learner language is important for building automated scoring systems and creating adaptive learning programs, including intelligent tutoring systems. The analysis of native language is useful for identifying the reading level of texts and using authentic language to generate exercises and other materials for language learners. As NLP continues to advance, its integration in language classes is expected to become more widespread, further enhancing the effectiveness of language teaching and learning processes.

Antoniadis and Desmet (2016) identified seven possible applications of NLP for language teaching: (1) a resource generator to create reference and teaching materials, such as learner

dictionaries, textbooks, and assessments; (2) a reading companion that can assist learners with second language reading comprehension by providing comprehension checks and definitions or translations; (3) an exercise and test generator that can adapt to learners' abilities based on the analysis of errors; (4) an error detector and automated essay scoring tool for assigning grades and providing feedback; (5) a writing aid and suggestion tool that supports learners in producing well-formed essays and other written responses; (6) an input provider that automatically selects comprehensible reading material based on learners' reading levels; and (7) an adaptive item sequencer that develops learning environments based on student input. An additional application is the creation of learning analytics that can provide general measures of language development (Chen, 2018; Kyle, 2016; Lu, 2014).

Table 1.1

Applications of NLP to Language Teaching

Analysis of Learner Language	Analysis of Native Language	Analysis of Both
Error detector AES Learning analytics	Resource generator Reading companion Activity/test generator	Input provider Adaptive intelligent tutoring systems

Automated Essay Scoring

AES applications date back as early as the 1960s with Ellis Page's classic paper on the possibility of AES (Klebanov & Madnani, 2022). Because NLP technology was not available at the time, Page's paper focused on questions of utility and need for AES, rather than on the actual tools and development. Page (1966) argued the need for AES was obvious; students need extensive feedback and correction on assignments, but most teachers do not have enough time, and perhaps are not even qualified, to offer such feedback. Due to what he called the

"fuzziness and inutility" of teachers' thinking, Page believed computers would offer higher quality large-scale essay scoring at a lower cost. He further argued that AES presented a feasible option, even though the technology did not exist at the time, because computers are smart and can learn from experience and modify behavior.

Page's arguments still hold true today. From a utility and needs perspective, colleges continue to report that today's students are not well-prepared for writing assignments and would benefit from more writing practice with extensive feedback (Burstein et al., 2016; Mezler, 2014). AES systems can reduce writing teachers' workload, allowing teachers to assign more essays and to focus their feedback on ideas and content rather than on surface features like grammar and mechanics. From a quality perspective, human-machine correlations have been proven to be very high (Attali & Burstein, 2004; Shermis & Hamner, 2013) and continue to improve as larger corpora become available. From a feasibility perspective, NLP techniques have allowed AES technology to catch up with the research, making Page's feasibility argument stronger than ever.

The evolution of AES followed the same pattern as the development of NLP in general. Early AES systems provided the basic scoring of texts focusing on grammatical and lexical correctness (Shermis & Burstein, 2013). Syntactic analysis was employed to detect and possibly correct writing errors. However, quality writing requires far more than simple grammatical correctness; it necessitates attention to organization, development of ideas, cohesion, linguistic sophistication, transitions, word choice, and coherence. With the development of LSA and AI techniques, AES systems have grown more sophisticated, now encompassing semantic, pragmatic, and discourse analysis. Semantic analysis measures the relevance of content in

relation to the writing prompt (Higgins, Burstein, & Attali, 2006), while pragmatic analysis assesses sociocultural aspects of language (Johnson, 2007). Discourse analysis can be used to assess cohesion and coherence (Miltsakaki & Kukich, 2004). As a result, contemporary AES systems are better equipped to analyze various facets of written text, providing more comprehensive evaluations of writing quality.

Researchers and developers are primarily concerned with building efficient, accurate systems and advancing the field by sharing new knowledge and tools. As AES systems improve and gain widespread traction, developers should remember AES is not solely an NLP enterprise but instead involves several stakeholders. Klebanov and Madnani (2022) identified other stakeholders as test-takers, institutions, teachers, subject-matter experts, and business units, and described the context and challenges for each group.

Test-takers depend on scores for admission, placement, or certification and require score reports to understand why they received their given scores and what areas they need to improve. Likewise, institutions rely on scores to make informed decisions regarding admission and placement, as well as to inform policy and funding choices. In terms of meeting the needs of test-takers and institutions, NLP developers face the challenge of balancing sophisticated, efficient models with the interpretability of scores, while ensuring that scoring remains free from bias.

Teachers are directly impacted by scores used to place students in classes. Teachers are also affected if the scores are used directly in the classroom for formative and summative assessment. Consequently, NLP developers need to create pedagogically sound systems with practical classroom applications that can enhance learning. Subject-matter experts develop the

writing prompts and scoring rubrics used to build AES systems, posing a challenge for developers to translate this expertise into valid, reliable instruments that can be efficiently computed.

Business units, often large technology companies with substantial financial resources, are responsible for developing and administering assessments. The challenge for developers lies in balancing the business-driven focus on profitability, speed-to-market, and cost reduction with the need to create accurate, reliable, and user-friendly AES systems that genuinely benefit test-takers, institutions, and teachers.

Table 1.2

Stakeholders Affecting NLP Research and Development

Stakeholder	Context	Challenges for developers
Test-takers	Obtain admission, placement, certification; improve weak testing areas	Produce results that are easy to interpret; ensure fairness
Institutions	Make admission and placement decisions; determine policy and funding	
Teachers	Teach students placed in classes; use results from formative and summative assessments	Build easy-to-use, pedagogically sound systems that have practical application and enhance learning
Subject-matter experts	Develop writing prompts and scoring rubrics used to build AES systems	Convert and translate test questions and rubric into valid, reliable system
Business units	Build, fund, and market systems; administer assessments	Get systems to market quickly; stay within budget

Meeting the multitude of demands from various stakeholders is no small task, but NLP developers must not lose sight of teachers and learners. Developers should work closely with teachers to determine how to build better systems that will reap the most educational benefits. Developer Vik Paruchuri (2013) acknowledged this need, arguing “AES is useless when the

power is in the hands of researchers and programmers. The real people who need to shape and implement these technologies are teachers and students, and they need the power to define how the AES looks and works." Paruchuri further argued that to be part of the development process and in a position to contribute, teachers need to understand what the code is doing. Those in ESL education stand to benefit greatly from such systems and therefore need to be included in subsequent work involving AES research. This requires cooperation on both the part of the developers and the teachers.

Contexts for Assessing Learner English

Writing instruction for non-native speakers (NNSs) of English varies depending on context. One major distinction lies between English as a Foreign Language (EFL) and English as a Second Language (ESL). EFL refers to English instruction in a country in which English is not an official language. ESL refers to English instruction within a native-speaking context. For example, NNS international students studying at American colleges or immigrant children enrolled in schools in English-speaking countries are ESL students. The native tongue of the teacher is irrelevant; both ESL and EFL teachers could be native or non-native speakers. The distinction lies in the environment, not the teacher. Whether the context is EFL or ESL, the students are referred to as English language learners, or ELLs. The written or spoken language produced by ELLs is referred to as learner English. The distinction between EFL and ESL affects the purpose of writing assessment.

Writing assessment in the EFL context usually measures linguistic proficiency, with a heavy focus on syntax and vocabulary. One of the main purposes of EFL writing assessment is to meet the application requirements for college students studying abroad. Most U.S. universities

require the Test of English as a Foreign Language (TOEFL), developed by Educational Testing Services (ETS). According to the ETS website (2022), the TOEFL iBT (internet-based test) is accepted at more than 11,500 higher education institutions in over 160 countries and has been administered to over 35 million test takers. The test consists of reading, listening, speaking, and writing sections. The testing organization explains, "The Speaking and Writing sections are scored by a combination of automated AI scoring and multiple, highly trained human raters to offer a complete and accurate picture of your writing ability, minimize rater bias, and ensure consistency and highest quality" (ETS.org, 2022). This statement suggests the use of AES in TOEFL scoring, which is not surprising given that ETS has been a leader in the development of AES. Another purpose of EFL writing assessment is for students to meet the English requirements of universities in their home countries. In China, for example, undergraduates must demonstrate English proficiency through the College English Test (CET), with writing and translation making up about a quarter of the total score (Zheng & Cheng, 2008). A third purpose is for workplace certification. Well-known assessments include two Cambridge exams, the First Certificate in English (FCE) and the Certificate in Advanced English (CAE), both of which include human-scored writing components (Ke & Ng, 2019).

Assessment in the ESL context may also aim to measure language proficiency, especially for students enrolled in ESL programs that separate students from native speakers. Many colleges and universities offer intensive English programs (IEPs) to provide support and language instruction with the goal of students transitioning from ESL to mainstream academic classes. Writing assessment is then used to determine whether students need additional ESL education and, if so, to place students in the appropriate level. The question for using AES in

this context is how many writing samples are needed to provide a sufficient measure in determining which language services and support students' needs to succeed academically. Uzen (2018) cites the creation of a sizeable corpus appropriate for training data as one of the main challenges in AES implementation. Another question concerns feature extraction. What writing features best identify a student's English proficiency? This question is complicated by the various language backgrounds of ELLs, with certain features being more representative of some languages than others, possibly leading to machine bias.

In addition to measuring language proficiency, writing assessment in the ESL context may also include measuring writing ability, assessing students alongside their English-speaking classmates. This occurs when ESL students are enrolled in mainstream classes, not separate programs, and are therefore assessed in the same manner as any other student. Unlike proficiency, writing *ability* focuses more on rhetoric, style, tone, development, cohesion, argument, and precision, rather than granular linguistic features like vocabulary and syntax. The question for using AES then centers around validity and fairness. With extensive exposure to English texts, native English speakers have linguistic intuition that allows for automatic control over grammar and vocabulary, freeing up mental energy to focus instead on idea generation and development, whereas NNSs vary greatly in their control over grammar and vocabulary and are forced to split their attention between language and content (Weigle, 2013). With that in mind, is it fair to put native and non-native speakers in the same playing field when it comes to implementing AES systems?

Benefits of Using AES in ESL Instruction

In both the EFL and ESL contexts, automated scoring has the potential to enhance

teaching and learning if the AES systems are implemented carefully and appropriately. Deane (2013) argued that the case for AES is strong when used to help students identify errors, practice writing, and improve their fluency. However, the case for AES is relatively weak when used to assess argument quality, rhetoric, tone, and other elements that differentiate students who already have a good command over the writing process. While much of the research on AES focuses on its use in large-scale, high stakes testing, such as the TOEFL, the intent of this study is to consider its use at the classroom level. In this regard, AES offers several benefits for ESL learners and teachers.

Error Correction

A prominent affordance of AES systems is the instant identification of lexical and syntactic errors. Error correction has long been a hot topic in ESL and EFL education. Early ESL teaching methods focused heavily on grammar. The audiolingual method popular in the 1950s, for example, drew on the behaviorist theory that learning occurs through conditioning and emphasized grammatical accuracy over fluency (Larsen-Freeman & Anderson, 2011). In the 1970s, ESL researchers began to criticize the focus on grammar, arguing that language is primarily social (Halliday, 1973), and that students may learn grammar rules and still be unable to communicate in the language (Wilkins, 1976). Hymes (1971) reasoned that beyond a simple understanding of linguistic rules, language acquisition requires communicative competence. From these observations, the communicative approach was born and, along with related approaches such as content-based instruction (Snow, 1991) and task-based learning (Ellis, 2003), it has become the dominant teaching method in ESL. With a new focus on the social and cultural aspects of communication, researchers began to criticize the use of grammar

correction. Krashen (1982) presented perhaps the most extreme view, arguing that grammar instruction is harmful to students and potentially hinders language acquisition. Other researchers claimed error correction to be simply ineffective in the development of language acquisition and writing accuracy (Kepner, 1991; Sheppard, 1992; Truscott, 2007).

While grammar correction in the U.S. was de-emphasized or even discouraged from the 1970s through the early 2000s, a shift began in the 2000s with a new focus on the role of grammar correction in ESL writing instruction. Ferris (1999) claimed that ESL teachers cannot simply ignore grammar mistakes and that clear and consistent correction will lead to improved writing. Several studies investigated the positive effects of indirect grammar feedback, in which teachers indicate errors but ask students to self-correct (Ashwell, 2000; Chandler, 2003; Ferris & Hedgcock, 2005). More recent studies have also found that focused corrective feedback can facilitate learning (Bitchener, 2008; Ellis et al., 2008; Sheen, 2007). Bitchener (2008) found that the combination of written corrective feedback and conferencing with ELLs significantly improves writing accuracy. In a study on the use of written corrective feedback, Sheen (2007) found that corrective feedback works best when it is intensive and targeted on specific errors. Sheen, Wright, and Moldawa (2009) expanded on Sheen's original study by conducting similar research and drawing these conclusions: "Focused CF [corrective feedback] may enhance learning by helping learners to (1) notice their errors in their written work, (2) engage in hypothesis testing in a systematic way, and (3) monitor the accuracy of their writing by tapping into their existing explicit grammatical knowledge" (p. 568).

Regardless of whether research findings indicate corrective feedback promotes language learning and increases accuracy, many ELLs *want* feedback on errors and believe it is

important to their linguistic growth. In a thorough review of research on corrective feedback on learner English, Ferris (2011) found studies consistently show students value comprehensive error correction from their teachers. However, due to time constraints, marking every single error on students' essays is not feasible for most ESL teachers, and so AES systems offer an obvious benefit. By identifying syntactic and lexical errors, using an AES system could free up time for teachers to focus on process writing and content development, both of which require teacher interpretation and are not easily automated.

Timeliness

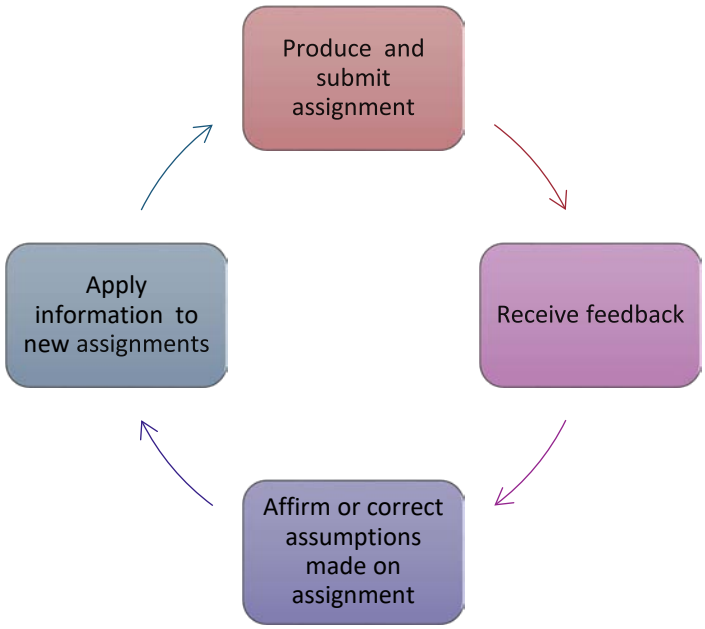
Decades of research on feedback in various learning contexts, not just writing instruction, consistently demonstrates that the sooner students receive feedback, the more effective it is (Hattie, 2008). One of the critical components of effective feedback is ensuring that students receive it when the information remains relevant and can assist them in upcoming assignments, enabling swift revisions and additional feedback opportunities (Li et al., 2015). This iterative process, illustrated in Figure 1.1, is referred to as the feedback loop (Hattie, 2008). If the feedback loop is broken, students begin each assignment anew and will likely repeat the same errors.

In the context of language learning, the application of newly acquired knowledge is particularly advantageous, as it facilitates the encoding of information by connecting it to prior knowledge. Providing timely feedback allows learners to actively engage with the learning material, reinforce their understanding, and address misconceptions or mistakes more effectively. Consequently, this enhances their ability to internalize linguistic structures,

vocabulary, and usage patterns, resulting in improved language proficiency and retention. The feedback loop, therefore, plays a crucial role in promoting continuous learning and progress.

Figure 1.1

Feedback Loop in Writing Instruction



Immediate feedback is ideal but impractical for teachers. Writing teachers often use peer review as a strategy for ensuring timely feedback on writing assignments (Wiggins, 2012). Though peer review is a proven technique in mainstream composition classes (Gere, 1987), it presents several problems in ESL classes (Leki, 1990). While peer review affords an excellent opportunity for developing social communication skills, studies have revealed that ESL students tend to focus on editing rather than responding to content, and since they are NNSs themselves, they may not offer correct editing advice, thereby rendering the feedback useless or even harmful. The use of AES may present a more effective strategy, as it makes possible the delivery of timely, meaningful feedback (Zhu et al., 2019). Lynch (2019) proposes the use of

automated assessments to consistently measure student growth and provide the timely, targeted feedback students can use to improve learning.

Consistency

In addition to time constraints, consistency also contributes to the value and efficacy of corrective feedback. Zamel (1985) criticized ESL teachers' feedback on student writing as being "confusing, arbitrary, and inaccessible" (p. 79). While Zamel was referring to the inconsistencies that may exist within an individual teacher's feedback system, the problem is further complicated when considering the huge variations in feedback from one teacher to another. Weigle (2013) attributes at least part of the problem to teacher training. Different ESL programs require different credentials. A linguistics degree requires coursework in syntax, but not as it applies to teaching. A degree in TESOL (teaching English to speakers of other languages) may focus on grammar instruction, but not necessarily in the context of teaching writing. A degree in English may provide insight in teaching writing to native speakers, but not ELLs.

Automated feedback can provide the consistency that students want and need, as long as it is presented in a way that is useful for learning. To prove useful, automated systems need to be able to identify the kinds of errors ELLs specifically are likely to make in contrast to general usage errors characteristic of both native and learner English (Weigle, 2013). Moreover, AES systems should identify the linguistic features that characterize different proficiency levels so that students can receive targeted, level-appropriate feedback that will help them progress. TESOL research indicates that the most important errors in ELL writing are those that occur frequently, interfere with comprehension, relate directly to the context of the assignment or teaching situation, and are at the right level so that students are developmentally ready to

understand and correct the error (Williams, 2005; Ferris, 2011, Weigle, 2013). AES systems can potentially provide students the feedback they need to self-learn and revise their own writing, leading to greater student autonomy (Ranalli et al., 2017).

Learning Analytics

Learning analytics (LA) has been defined as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environment in which it occurs" (Long & Siemens, 2011, p. 34). In other words, LA uses data generated from students for the prediction of educational outcomes, with the goal of supporting students and enhancing their education. An AES system could be a powerful LA tool for informing ESL instruction. At the individual student level, teachers could use the data to create a personalized learning plan with specific, tailored goals. At the classroom level, teachers could examine the data generated from a whole class to determine which curricular objectives to prioritize. Beyond the classroom, English language teachers, program administrators, and materials writers could use AES analytics to inform curriculum planning and decisions.

Challenges with AES Implementation

Despite the numerous benefits of AES, its implementation is not without challenges. Several factors contribute to the relatively limited adoption of AES systems in English language teaching. From a student perspective standpoint, some studies indicate leaners may feel less motivated to write when their audience is a machine instead of a human, often expressing doubts or objections to scores assigned by a machine that has no understanding of the ideas or

concepts contained in an essay (Barker, 2011; Lai, 2010; Landauer et al., 2000; Weigle, 2013). For language programs and teachers, challenges include the cost of implementation, teacher resistance to learning and implementing a new technology, and concerns regarding whether the technology aligns with their specific needs. Addressing these barriers is essential to fostering greater acceptance and integration of AES systems in English language teaching.

Cost

A major concern regarding AES tools is cost. Over time, AES systems should be far less expensive than human graders, yet school and program administrators may be reluctant to make the initial investment, especially considering that technological tools can become dated quickly, requiring further costly updates or replacement (Gartner Inc., 2021). While several open-source or other freely available systems exist, they are often in an early development stage and do not offer sufficient accuracy for widespread adoption by an ESL program. Not surprisingly, once an AES system achieves a high level of accuracy, it becomes a prized intellectual property that is offered only as a propriety solution at a high-cost point (Kumar et al., 2017).

Sophisticated AES systems may prove to be particularly cost-prohibitive to ESL programs, which have notoriously tight budgets. In higher education settings, ESL programs tend to be treated as ancillary to academic programs and are not given the same resources as mainstream academic departments (Osborne, 2015). The financial situation is not much better in public K-12 institutions. Williams (2020) reported that although English language learners make up one of the fastest-growing student populations in the U.S., funding for ESL programs has remained flat since 2002.

Teacher Resistance

Teacher buy-in presents another challenge. ESL teachers may be hesitant to integrate AES in the classroom due to a fear that automated systems will replace human instruction. In a perception study on foreign language instructors at the Ohio State University, Amaral (2011) found that many foreign language teachers view digital tools as distractions that take time away from human-to-human communicative activities. Language learning relies on human interaction for students to develop the necessary skills required for negotiating meaning, understanding sociocultural behavior, and observing nonlinguistic communication such as body language and facial expressions. That said, teachers expressed interest in the use of automated tools to support, not replace, human instruction (Amaral & Meurers, 2011). This should always be the goal of educational technology; digital tools are designed to enhance learning, not replace high-quality teaching. Because they focus on text production features such as grammatical accuracy and cohesion rather than on critical thinking and creativity, AES systems particularly are not suited nor intended to replace teachers.

Researcher-Teacher Disconnect

A third explanation is the disconnect between developers and practitioners: "The development of systems using NLP technology is not on the agenda of most CALL [computer-assisted language learning] experts, and interdisciplinary research projects integrating computational linguists and foreign language teachers remain very rare" (Amaral & Meurers, 2011). Computational linguists who develop AES systems may be unaware of teaching and learning issues such as second language acquisition models, current teaching methods, language policy and planning, curriculum and activity design, and language assessment.

However, the past decade has seen a great increase in interdisciplinary research between NLP and language teaching (Fu, Gu, & Yang, 2020). The COVID-19 pandemic further accelerated the widespread use of NLP-based digital learning tools for educational purposes in general (Almarzooq, Lopes, & Kochar, 2020), leading to even more research projects and opportunities between AES developers and teaching practitioners.

Solutions to the Challenges

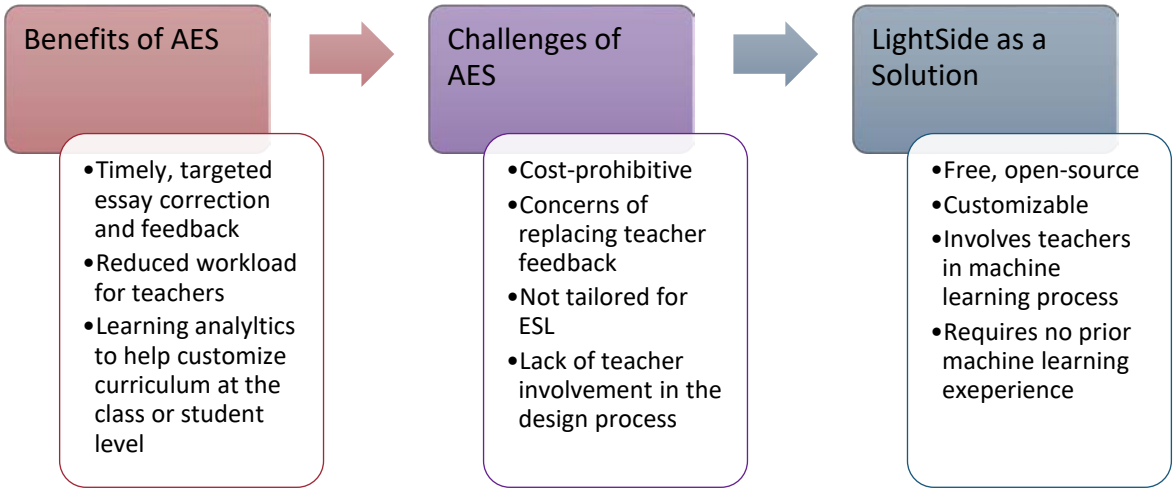
Studies reveal that teacher failure to implement educational technology is due more often to teacher resistance than to any shortcomings of the technology itself (Curan, 2005; Grimes & Warschauer, 2010; Weigle, 2013). Citing administrative support, professional development opportunities, and teachers' readiness to experiment as key factors in lessening teacher resistance, Weigle (2013) suggests, "One possible path is to expose teachers to non-commercially produced automated tools that can be used to explore dimensions of their students' writing" (p. 75).

LightSide (Light Summarization Integrated Development Environment) offers such a tool. Created by researchers at the Language Technology Institute at Carnegie Mellon University, LightSide is a machine learning text-mining platform that streamlines the process of extracting features from a text to create a predictive model that assigns scores to essays. As an open-source technology, LightSide is free for anyone to download and use. The LightSide user's manual promises to make machine learning easy for anyone: "We've built a tool that lets you hit the ground running with your data, putting as much of the research workflow for machine learning as possible into an easy, point-and-click interface" (p. 1). Most importantly, LightSide produces excellent results. In a comparative study on AES systems, LightSide performed as well

as eight proprietary systems developed by large-scale testing companies such as ETS (Shermis & Hamner, 2012). Figure 1.2 summarizes the benefits and challenges of AES and considers LightSide as a potential solution.

Figure 1.2

Benefits and Challenges of AES in ESL Instruction



CHAPTER 2

RELATED LITERATURE

Automated essay scoring (AES) has attracted great attention in educational research, though much of that attention focuses on K-16 education in the U.S., with little or no special consideration given to students whose native language is not English (Hibert, 2019). One of the most promising benefits of AES is the quick delivery of consistent scores, free of human error such as bias, grading fatigue, varying levels of rater expertise, perception differences, halo effect, and distractions (Kumar & Boulanger, 2020). AES, however, presents its own set of challenges such as imperfections in training data, over- or underrepresented minority groups, and questionable correlations, all of which are amplified in ESL, as research in the field, especially as it applies to lower-proficiency language learners, trails research on native-speaker English (Huang & Renandya, 2020).

While research on AES systems dates to the 1960s, AES studies specific to the ESL/EFL context did not appear until the late 1990s (Dikli, 2010). Burstein and Chowdorow (1999) examined the performance of e-rater on the Test of Written English (TWE), comparing a sample of essays written by English language learners to a sample of essays by native English-speakers. The results showed significant differences between the scores of the two groups, pointing to the need for further research on the topic. The past two decades have witnessed a growing body of research on reliability and validity, usage, impact, and perceptions of AES as it relates to ESL teaching and learning. This chapter presents a broad overview of that literature and a more focused look at research pertaining specifically to the use of LightSide in writing instruction. The review seeks to answer the following questions:

1. What major AES systems and learner (non-native) corpora are currently available?
2. How have AES systems been tested for accuracy?
3. How have AES systems been used to facilitate language learning?
4. What are the major challenges in using AES in language education?
5. What are teachers' perceptions of using AES in language classes?
6. How has LightSide been used as an AES system?

Currently Available AES Systems and Learner Corpora

Before discussing usage and perception studies of AES in ESL, an overview of some of the most well-known AES systems and learner corpora provides a foundation for understanding the topic. AES, also referred to as AEG (automated essay grading) or AWE (automated writing evaluation), systems include both commercial and non-commercial tools intended for the evaluation of written language. These systems are trained with corpora, large digital collections of texts. Learner corpora refers to a collection of texts written by English language learners and are essential for training AES systems to evaluate learner English.

AES Systems

A variety of AES systems are available. Although these systems were originally developed to assess native English, many have added special features, including multilingual feedback systems, that are useful for evaluating learner English (Warschauer & Ware, 2006). AES scoring strategies may be holistic or trait-based. A holistic score measures and weighs evaluative criteria to produce a single, final score based on the overall quality of the assignment. Trait-based scoring provides separate scores for different dimensions of writing, such as grammar, vocabulary, mechanics, and organization (Shermis & Burstein, 2013). In

addition to scores, some systems provide detailed feedback and other supporting features such as rubrics and model essays (Warschauer & Grimes, 2008). The following represent some of the most well-known AES systems.

Bookette

Developed by CTB/McGraw Hill, Bookette uses NLP with a neural network to model human scores by training the engine using human-scored essays and validating the engine against a separate set of human-scored essays (Shermis & Hamner, 2012). CTB has developed AES systems for large-scale testing since 2009 and for classroom settings since 2005 (Shermis & Hamner, 2012). It builds both prompt-specific and generic scoring engines, with the prompt-specific engines providing greater reliability. The engines classify around 90 text features and categorize them according to the traits of organization, development, sentence structure, word choice, grammar usage, and mechanics. The trait level scores can be reported separately or combined as a single score (Shermis & Hamner, 2012).

e-rater and Criterion

Educational Testing Service (ETS), the largest nonprofit educational assessment organization in the world, created e-rater in 1998. e-rater uses statistical, rule-based NLP methods to predict a holistic score based on a grading rubric that measures grammar, mechanics, vocabulary, style, and essay development (Burstein et al., 2013). To achieve this, e-rater extracts eleven features in two categories, content and writing quality (Hussein et al., 2019). Writing quality features include grammar, usage, word choice, word length, mechanics, development, style, and organization, while content features rely on prompt-specific

vocabulary (Ramineni & Williamson, 2018). Features are extracted and analyzed in a training sample of 250 or more human-scored samples and weighted using a regression model approach to provide a sum of the weighted features that is then calculated to predict a final holistic score (Burstein et al., 2013).

While e-rater is best suited for placement or summative assessment, ETS's other AES platform, Criterion, is designed as a formative feedback tool. Criterion includes a single holistic score but also provides detailed feedback on grammar, mechanics, vocabulary, style, and organization (Burstein et al., 2013). Criterion is an online tool designed to be used as a platform for instructors to grade and provide feedback to student writing submissions (Song, 2012). Criterion generates a holistic score and provides automated feedback on grammar and mechanics. Though a holistic score is provided, the intended use of Criterion is as a supplementary classroom tool under an instructor's supervision, since the tool does not evaluate content or logic (Song, 2012).

Intelligent Essay Assessor and WriteToLearn

Landauer (2003) developed Intelligent Essay Assessor (IEA) for Knowledge Analysis Technologies. The system was later acquired by Pearson Knowledge Technologies. Intelligent Essay Assessor uses latent semantic analysis (LSA) to analyze around 60 variables that contribute to a total essay score (Foltz et al., 2013). It can be trained to evaluate text in any language, not just English (Shermis & Hamner, 2012). IEA can be trained using prompt-specific algorithms to match student essays to human scores and can detect off-topic responses. It provides an immediate evaluation of an essay and includes trait-specific feedback on errors in grammar and mechanics. The system can also detect plagiarism, an important feature that is

difficult for human graders to spot but is very important for addressing academic dishonesty (Dikli, 2006). Unlike most other AES systems, which require 300-500 essay samples to train each prompt, IEA requires only 100 prescored training samples (Hussein et al., 2019).

Intelligent Essay Assessor was used as the basis for Pearson's WriteToLearn program, a web-based platform that provides learners with essay prompts and text summary activities. WriteToLearn is designed as a formative tool that provides continuous assessment and detailed feedback on trait features, similar to ETS's Criterion (Shermis & Hamner, 2012). While IEA is designed for use by institutions, WriteToLearn is intended for student use, either as a self-study tool or part of classroom instruction (pearsonassessments.com).

IntelligMetric and My Access!

Like ETS and Pearson Knowledge Technologies, Vantage Learning offers two AES tools. The first, IntelliMetric, provides a holistic score based on features of grammar, mechanics, syntactic complexity, support, and cohesion (Schultz, 2013). IntelliMetric was developed using principles from NLP, latent semantic analysis, AI, and natural language processing (Shermis & Hamner, 2012). Developed in 1998, it is regarded as the first AES system that uses artificial intelligence to imitate the manual scoring process of human raters (Hussein et al., 2019). It identifies text characteristics as larger categories referred to as Latent Semantic Dimensions. It can score essays in several languages other than English, including Arabic, Dutch, French, German, Hebrew, Italian, Japanese, Portuguese, and Spanish (Elliot, 2003).

The second tool, My Access!, is a writing aid, much like Criterion and WriteToLearn, that provides both a total score and detailed feedback on trait features, including focus and meaning, content and development, language use voice and style, and mechanics and

conventions (Vitartas et al., 2016). It offers more than 200 writing prompts in several genres like narrative, informative, and persuasive writing (Hussein et al., 2019). Like its ETS and Pearson counterparts, My Access! could potentially be used for self-study but is intended for use as an augmenting tool for teachers.

LightSide and Revision Assistant

A free open-source package developed at Carnegie Mellon University, LightSide (Light Summarization Integrated Development Environment) is not a polished, ready-to-use AES system, but rather a tool that allows non-experts to perform text-mining for a wide range of purposes, including essay scoring. Predictive models can be trained and tuned by using the standard options available on the user interface (Shermis & Hamner, 2012). While LightSide includes several plugins that allow for more options in machine learning, data representation, and visualization, predictive models can be trained and tuned using only its standard options. The system is unique in that it is open-source and allows practitioners to take on the role of researcher by training and testing systems using their own data sets.

LightSide partnered with Turnitin, a well-known plagiarism checker used in many schools, to create "Revision Assistant," an online tool that highlights passages from student writing and offers suggestions for improvement based on the writing prompts and scoring rubrics uploaded by the teacher. It should also be noted that Turnitin partnered with ETS e-rater to provide a grammar checker that helps teachers quickly identify errors.

Project Essay Grade (PEG)

The scoring system developed by Ellis Page in 1966, PEG is often cited as the earliest AES

system (Dikli, 2006). As such, it has gone through decades of research, development, and improvement (Shermis & Hamner, 2012). Measurement Inc., the company which acquired PEG in 2002, claims on its website that PEG is the most researched AES system, has been used to provide more than two million scores to students, and is currently in use at 1,000 schools and 3,000 public libraries. The company makes the following claim: “Using advanced, proven statistical techniques, PEG analyzes written prose, calculates more than 300 measures that reflect the intrinsic characteristics of writing and achieves results that are comparable to human scorers in terms of reliability and validity” (measurementicn.com/peg). PEG uses correlation coefficients to predict scores. The system defines “trins” as intrinsic variables such as punctuation or grammar, and “proxes” as correlations between intrinsic variables like text length or average word length (Dikli, 2006; Hussein et al., 2019; Valenti et al., 2017). Scoring occurs in two stages. First, training sets are analyzed across several dimension, and features are extracted to measure syntax, mechanics, semantics, and organization. Second, PEG builds a predictive model to assign either holistic or trait-based scores (Shermis & Hamner, 2012). Training samples should consist of 100 to 400 essays. The training output produces a set of coefficients from the proxy variables. To score new essays, proxes are identified and used in the prediction equation, with a final score determined from the estimation of coefficients from the training samples (Dikli, 2006).

Critics claim PEG focuses too much on surface features while disregarding the semantic aspects of essays (Hussein et al., 2019). After its acquisition, Measurement Inc. has attempted to address this criticism. The company has continued to develop PEG on a deeper semantic level, adding new features to measure fluency, diction, and construction. Measurement Inc.

also created a custom search language that allows complex structures to be located within a text quickly (Hussein et al., 2019).

Table 2.1

Overview of AES Systems

System	Developer	Scoring Strategy	Instructional Application
Bookette	CTB/McGraw Hill	holistic and trait-based	N/A
e-rater	ETS	holistic and trait-based	Criterion
IEA	Knowledge Analysis Technologies	holistic and trait-based	WriteToLearn
IntelliMetric	Vantage Learning	holistic and trait-based	My Access!
LightSide	Carnegie Mellon	holistic	Revision Assistant
PEG	Ellis Page	holistic	N/A

Learner Corpora

Training and tuning AES systems requires corpora of student writing. A corpus is a principled collection of spoken or written language compiled in a database for the purpose of linguistic analysis (Biber et al., 1998; Sinclair, 2004). AES systems score essays by measuring and aggregating text features through computer algorithms to predict a score similar to what a human rater would assign (Dikli, 2006). The features are usually extracted from a corpus of essays scored by human raters. Since most AES systems were designed for native English assessment, they were trained and tested with corpora consisting of writing samples by native speakers, or at least with no distinction made between native and non-native writing. Shermis and Hamer (2012), for example, wrote an extensive comparison of the most popular AES systems based the Automated Student Assessment Prize (ASAP) corpus released as part of a Kaggle data analysis competition. Consisting of thousands of student essays, this corpus has

become widely used for holistic scoring and AES reliability and validity studies (Ke & Ng, 2019). However, building and training AES systems that can analyze learner English requires learner corpora. This section presents examples of publicly available learner corpora.

The Cambridge Learner Corpus – First Certificate in English

The CLC FCE dataset contains 1244 essays written by Cambridge First Certificate test takers in 2000-2001 and scored by professional graders (Yannakoudakis et al., 2011). The test takers responded to one of ten possible writing prompts. The dataset includes the original essay and a holistic score, in addition to marks, rater comments, error annotation, and demographic information including the test taker's age and first language. The data set represents test takers from 138 different first language backgrounds. It includes manually tagged linguistic errors, which are useful for building systems that can detect and correct grammatical scores, but the small number of essays provided for each writing prompt makes it hard to build prompt-specific systems (Ke & Ng, 2019).

The TOEFL 11 Corpus

Developed by ETS, the TOEFL 11 corpus contains 12,100 essays responding to eight prompts written by test takers for the TOEFL, an English placement exam for international students applying to US colleges and universities (Blanchard et al, 2013). The original dataset from 2006 and 2007 contained 1,100 essays by test takers with a linguistic background in Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. The essays in this corpus are available as both raw and tokenized forms. In addition to native language, the essays indicate the test taker's English proficiency level. In July 2014, new essays

were added to the corpus, bringing the collection to 12,100. The corpus offers only three levels of proficiency and can thus be used to train AES systems only on holistic scores.

The International Corpus of Learner English

The original ICLE contained 1003 essays corresponding to thirteen prompts, with 2.5 million words written by test takers from eleven different language backgrounds (Granger et al., 2009). The second version was released in 2009, adding 3.7 million words with test takers from sixteen language backgrounds. The second version included 830 essays responding to 13 new prompts. Developed for the purpose of dimension-specific scoring, the essays are annotated for essay quality, including dimensions of organization, thesis clarity, prompt adherence, and argument persuasiveness (Ke & Ng, 2019).

Table 2.2

Overview of Learner Corpora

Corpus	Number of Essays	Number of Languages	Number of Prompts	Scoring Range
CLC FCE	1244	138	10	1-40
TOEFL 11	12,100	11	8	low, medium, high
ICLE	1,003	11	13	1-4
	830	16	13	

Accuracy of AES Systems

Teachers' assumptions about the purpose and usefulness of AES systems are guided by the accuracy of such systems. An argument in support of classroom implementation of AES rests on the assumption that AES systems are accurate. As with all assessment scoring, AES applications must be validated in terms of their intended use and interpretation.

Reliability and Validity Studies

Reliability in writing assessment refers to the consistency of scores when test takers are assessed on different occasions or on different tasks (Huang, 2008). Factors such as testing location, temperature, time of day, clarity of instructions, writing topic, and distractions, may affect test takers' performance and therefore affect reliability. A reliable test minimizes the impact of such factors. Raters also contribute to reliability. Intra-rater reliability refers to the ability of one rater to apply consistent standards and scores to all test responses, while inter-rater reliability involves consistent scoring practices among different raters (Johnson et al., 2009). Validity, which refers to the accuracy on score interpretations, depends on reliability (Huang, 2008). Studies on the reliability and validity of the scores produced by AES systems tend to focus on the agreement rates between human and machine raters. Several studies compare the correlations between the scores assigned by AES systems and human raters to the correlations between scores assigned by two or more human raters, with the assumption that human-assigned scores are valid enough to accept as the "gold standard" (Powers et al., 2000). These studies have shown that the correlations between human raters and AES systems are approximately as high as the correlations between two human raters (Attali & Burstein, 2006; Shermis, 2014).

Reliability Challenges

Perhaps the first issue to consider regarding reliability is the very idea of human raters creating the "gold standard." Raczynski and Cohen (2018) distinguish two groups: raters and experts. Raters, who are brought onto a project solely for the purpose of assigning scores, typically undergo training and calibration practice consisting of several hours or days, whereas

experts are highly experienced, proven raters responsible for training new raters, monitoring raters' performance, and developing training and calibration materials (Raczynski et al., 2015). Studies have indicated raters are prone to inaccuracy and reliability issues such as severity or leniency, fatigue, distraction, and rater drift (Leckie & Baird, 2011; Raczynski & Cohen, 2018; Weigle, 2013). Raczynski and Cohen (2018) argue that experts, not raters, should be the real "gold standard," but suggest the reason raters are used more often is due to costs and possibly the postulation that rating errors are negated when many raters are used for generating the scores used to train and test AES systems. Rater reliability for non-native English writing raises many questions, including whether the raters have been trained specifically for the evaluation of learner English.

Approaches to Validity

Agreement between human and automated scores is valuable but may be insufficient to serve as the only indicator of AES reliability and validity. Agreement results fail to provide enough information about the construct validity of AES systems. A construct is a theoretical concept or abstract idea such as logical reasoning, critical thinking, or creativity; construct validity refers to the ability of an assessment to measure the concepts it claims to evaluate (Bhandari, 2022). Construct validity is a tricky issue particularly in ESL assessment because research on second language writing indicates language proficiency and writing ability are two separate constructs (Cumming, 1989; Sasaki & Hirose, 1996; Weigle, 2010). Yang et al. (2002) proposed three different approaches to validation; the first approach is the common practice of correlating human and automating scores, the second involves the comparison between

automated scores and external measures of a similar construct, and the third focuses on scoring processes.

There are a multitude of studies using the first approach of comparing automated and human scores, generally with results indicating automated systems can perform on par with human raters (Attali & Burstein, 2006; Bridgeman et al., 2012; Shermis, Burtein, Higgins, & Zechner, 2010). Identifying a gap in research that compares results across multiple platforms, Shermis and Hamner (2013) conducted an extensive study that compared nine popular AES engines on the basis of scores from independent raters. The engines included AutoScore, LightSide, Bookette, e-rater, Lexile Writing Analyzer, PEG, IEA, CRASE, and Intellimetric. The study involved 22,029 student essays written by students grades 7, 8, and 10 in U.S. public schools. The essays included persuasive, expository, and narrative writing with both source-based and non-source-based tasks. Scoring rubrics included both trait and holistic scoring strategies. The study evaluated scoring performance based on distributional differences (correspondence in mean and variance of human scores to AES scores) and agreement (measured by correlation, weighted kappa, and percent agreement). The results showed that most of the AES predictions were at least 90% accurate, as compared to the human scores.

An example of a study using the second approach, looking at external measures, can be seen in a study by Attali (2007), using a multitrait-multimethod approach, analyzing the essays of 5,006 TOEFL examinees from 31 countries who had repeated a writing test twice. After a correlational analysis was applied between all scores from the two tests to determine alternate-form reliability, the correlations of the essay scores assigned by a human rater and machine were analyzed together with TOEFL subscores (structured writing, reading, and

listening). The essay score correlations were then compared to essay length. As in the Attali and Burstein (2006) study, this analysis found autoscoring reliability (0.71) to be higher than both single human rater reliability (.54) and double human rater reliability (0.63). The reliability of TOEFL subscores was around 0.80.

The third approach, focusing on scoring processes, considers *meaningfulness* of the features and not just data-driven statistics (Yang et al., 2002). As an example, Landauer et al. (2001) compared the relative contributions of the different scoring components of the IEA and found the component with the largest contribution was related to content. Looking at the scoring process also involves analyzing the differences between human and AES scores in terms of features and the weighting of features (Yang et al., 2002). In other words, humans and AES systems may treat certain features, especially those related to content and meaning, differently when it comes to scoring essays, and those differences may explain inconsistencies or outliers. This approach is not common; in fact, it may seem counterintuitive and almost backwards, as it focuses on disagreement instead of agreement. A study by Powers et al. (2001) took this approach by asking experts to trick the AES into assigning higher scores than a writing sample deserved. This study is discussed in more detail in the section on system gaming.

Validity Challenges

Though AES validity studies have yielded overwhelmingly positive results, indicating that AES systems meet or even exceed the performance of human raters, such studies are designed for the general demonstration of the capacity of AES and do not suggest the systems are perfect, nor should the systems be considered as replacement for human raters (Shermis &

Hamner, 2013). Several important questions remain, and study limitations present serious challenges in validity.

Weak Agreement

Not all validity studies indicated strong agreement. In fact, several studies investigating the use of automated scoring in classroom-based writing assessment have shown lower agreement between automated and human scores compared to agreement among human raters (Bridgeman et al., 2009; Ebyary & Windeat, 2010; Hoang & Kunnan, 2016; Huang, 2014; James, 2006; Li et al., 2014; Liu & Kunnan, 2016). A study conducted by James (2006), for example, examined the accuracy of AES for placement testing at the University Preparation Department of a post-secondary program. Multiple writing samples were collected from 60 student participants. The essays were scored by Intellimetric and eleven untrained human raters from the same university. The results showed positive correlations between the scores assigned by human raters (between 0.45 and 0.80) but lower correlations between human rater scores and automated scores (between 0.40 and 0.61).

Wang and Brown (2007) investigated the validity and usefulness of AES in scoring different dimensions of essays in large-scale placement tests through a correlational study examining the holistic scores assigned by human raters and by the AES system, Intellimetric, on essays written by 107 Hispanic students for the WritePlace Plus test. The results showed no significant correlation. The researchers concluded that human scores and machine scores were consistent only in sentence structure and opposed the findings of Vantage Learning (2000), which claimed high consistency for content, organization, and style.

Bridgeman et al. (2009) conducted a study on the essays written by the eleventh-grade

students for an exit exam. The essays were assigned holistic scores by two human raters and e-rater. The research team found 0.84 agreement for human-human but only 0.76 agreement for human-machine scoring. Similarly, Hoang and Kunnan (2016) analyzed the agreement between scores assigned by My Access and human raters on ESL students' responses to different writing prompts. Their analysis revealed stronger agreement between the two human raters (0.78) than between human raters and My Access (0.68). In contrast, Liu and Kunnan (2016) found WriteToLearn produced more consistent but also more severe scores compared to human raters.

In their report, Shermis and Hamner (2013) concluded that agreement with human scores may not always be the best or the only measure of writing proficiency. They suggest supplementing with other measures such as alternate-form reliabilities and correlations with course grades. As noted earlier, the predictive model common in AES studies often fail to address construct validity. Moreover, Shermis and Hamner acknowledged some odd statistical properties and conflicts in documented correlation procedures for some tasks of their study.

System Gaming and Testing Washback

A major concern regarding the validity of automated scoring engines is that they can be "gamed," or fooled into assigning high scores to poorly written essays. Les Perelman, together with students from MIT and Harvard, created a gibberish-generating engine he called Babel (Basic Automatic B.S. Essay Language Generator). Babel generates essays based on up to three keywords, which are nonsensical to the human reader, but which he has shown receive high scores from several AES systems (Kolowich, 2014). However, this issue of gaming may be overcome by implementing a framework that Higgins and Heilman (2014) have developed to

quantify the level of a scoring engine's susceptibility to gaming strategies. The framework relies on current knowledge and hypotheses regarding gaming strategies and simulates those strategies using computational methods so that the scoring engine can identify test takers' attempts to fool the system.

A study by Powers et al. (2002) tested the threat of system-gaming by studying the sensitivity of AES systems to the extraneous features of test takers' writing skills. The research team invited 27 writing experts to trick e-rater into assigning scores higher or lower than they deserved. The experts were asked to write two complementary essays in response to two GRE writing prompts, with a total of four essays per expert. Their essays were then scored by e-rater and two human raters. The difference in scores was intended to indicate how and to what extent e-rater could be deceived. The results showed e-rater was vulnerable to the experts' tricks, with the experts consistently obtaining higher scores than the human raters felt they deserved. These findings suggest that automated scoring systems should always be used together with human scoring in high-stakes assessment contexts in order to deal with anomalies.

Similar to test-takers learning to game an assessment system is the notion of testing "washback," a term used to describe the influence of an assessment on the teaching and learning that occurs in preparation for that assessment (Green, 2020). In other words, washback could occur if ESL teachers base instruction primarily on a particular assessment, such as the TOEFL, rather than teaching with the general goal of helping students achieve better language proficiency. Validity studies, therefore, should consider how assessment

measures could be subject to manipulation by test-takers, and perhaps even teachers, who have an interest in boosting test scores.

Differences in Learner English

As mentioned earlier, automated scoring systems were originally designed to evaluate native English; thus, studies testing reliability and validity specifically for learner English provide valuable insight for ESL educators, yet research in this area remain somewhat scarce. In addition to showing AES scoring differences between native and non-native English speakers, Burstein and Chodorow's (1999) study also showed native language background affected human and machine scoring in difference ways, possibly indicating linguistic bias. Arabic and Spanish speakers, for example, got slightly higher scores from human raters than from e-rater, whereas essays of Chinese speakers got higher scores from e-rater, with a standard deviation of 0.48. In another study on ELLs, Chodorow and Burstein (2004) investigated the effect of essay length on the automated scores assigned to TOEFL essays. Using a mixed model repeated measures ANOVA, the researchers analyzed 265 training essays for each of seven prompts. Results indicated e-rater and human scores differed across language groups for only one prompt. Arabic and Japanese speakers received higher scores from e-rater, while Spanish speakers received similar scores both from human rater and e-rater. Essay length was found to be a factor.

An important aspect of evaluating AES systems is to consider overall fairness and whether subpopulations of test-takers are treated differently. This type of evaluation needs to apply both to differences between native and non-native English speakers, and to differences between language groups within the non-native category.

Summary of Accuracy Studies

Correlation studies indicate AES has developed sufficiently for use in low-stakes assessment, such as formative assessment measures in classes, and as a second scorer for high-stakes assessment (Shermis & Hamner, 2013). However, practitioners should proceed with caution and be aware of the many challenges affecting reliability and validity. These studies underscore the notion that AES systems are a useful tool for supplementing, not replacing, human raters. Furthermore, Ranalli et al. (2017) claim AES accuracy studies are more developer-centric rather than user-centric and tend to overlook the question of how accuracy affects usefulness, concluding that more research is needed, especially in ESL, to ensure AES systems are accurate enough to provide pedagogical value.

Usage Studies

The use of AES systems as educational tools in ESL and EFL contexts is gaining traction (Stevenson & Phaktiti, 2014). Several studies have explored the use of AES to assist ELLs in the revision process of essay writing. Rock (2007) examined the impact of short-term use of Criterion on students' writing scores, concluding that students who used Criterion for supplemental writing instruction received higher scores and significantly improved the mechanical aspects of writing. Ebyary and Windeat (2010) found students made significant progress from rough draft to final draft, with four drafts in total, when provided with regular feedback from Criterion. A similar study examining the impact of EFL students using My Access! between drafts found that students in the treatment group felt motivated to revise more and write longer essays (Chou et al., 2016). The research team concluded using My Access! improved students' writing quality. In another EFL study, Tang and Rich (2017) examined the

impact of using an AES system throughout the writing process, with findings that students in the treatment group made greater progress and received higher scores on their final drafts than students in the control group. Li et al. (2014) investigated the use of Criterion for classroom-based formative assessment and found Criterion to be a helpful tool in motivating students throughout the writing process.

Some usage studies focus specifically on the impact of AES on error reduction. In a study investigating how students use Criterion feedback to correct their errors, Attali (2004) found students receiving regular feedback from Criterion reduced their errors in grammar, usage, mechanics, and style. Similarly, the results of studies by Kellogg et al. (2010) and Liao (2015) on the impact of Criterion feedback showed students reduced errors in grammar and mechanics in both revisions and new assignments. In an EFL study on the effects the CorrectEnglish, an AES developed specifically for ELLs, Wang et al. (2013) found students using the program reduced the number of errors on their writing assignments. Gao and Ma (2019) studied with effect of automated corrective feedback on ELLs ability to correct errors. Their study specifically focused on errors and corrections about past tense verbs. The treatment group that received automated feedback made greater progress in their ability to detect errors than the control group. However, the two groups did not differ in their ability to transfer what they learned to subsequent writing tasks.

Another use of AES systems is the development of autocompletion tools. Instead of correcting errors made by students, autocompletion proactively offers suggestions to assist students as they are typing. Yen et al. (2015) used text mining to identify grammar patterns and create WriteAhead, a program that assists ELLs by suggesting words and phrases to complete a

given sentence. WriteAhead organizes, summarizes, and ranks suggestions so that students can make an informed decision when choosing what to write (Yen et al., 2015). A research team in Taiwan took autocompletion in a new direction by developing RESOLVE, a context-aware emotion synonym suggestion system that applies sentiment analysis to language learning (Chen et al., 2018). The results of their study showed participants made significant progress in having a better command of emotion words and connotations in their writing.

Challenges in Usage

The use of AES in education raises several pressing concerns. As mentioned previously, test takers' ability to game the system is a serious challenge. Somewhat related is the concern that test takers will focus only on the final product rather than the process of writing (Dikli, 2020). These questions arise in writing instruction in general, independent from digital environments. Composition teachers have long criticized the prescriptive five-paragraph essay model that boils the process of essay composition into a series of efficient tasks which make writing more about completing a template than thinking critically, generating ideas, and artfully expressing those ideas (Warner, 2018). One concern specific to technology is that AES systems are vulnerable to irrelevant responses that are well-constructed but do not directly answer a question or appropriately respond to a writing prompt (Horbach & Zesch, 2019).

Perception Studies

The increased use of AES has sparked controversy, with teachers expressing doubt at the ability of a machine to evaluate writing and students balking at the idea of composing texts for a machine rather than a human audience (Weigle, 2013). To combat negative perceptions, it

is necessary to emphasize the use of AES as a supplementary tool capable of providing students with immediate feedback and allowing teachers to focus on the more substantive elements of writing such as style and the development of ideas instead of superficial features like grammar and mechanics. Recent perception studies have revealed growing trust in using AES to augment ESL teaching and learning.

In one study on teachers' perceptions of using an AES system as a tool to help teachers with classroom management by saving time and allowing teachers to focus more on higher-level writing concerns such as content and development, rather than superficial issues like grammar and mechanics, teachers indicated that AES was useful and created a more pleasant teaching experience (Grimes & Warschauer, 2010). However, in that same study, teachers showed low confidence in the scoring and preferred balancing AES with traditional teaching methods. The researchers found successful implementation is affected by teachers' familiarity with the technology.

A small-scale study of English as a Second Language (ESL) students in a pre-university writing course at Iowa State University found that about half of the feedback provided by Criterion was disregarded by students, who had lost trust in the system after finding inaccuracies in some of the feedback, such as identifying proper nouns as spelling errors, or correct sentences as fragments or run-ons (Chapelle et al. 2015). Even so, Chapelle et al. (2015) concluded, "Given that the proportion of successful revision is over 70%, Criterion® feedback can be considered as positively influencing the revision process, even if substantial room for improvement exists."

The Conference on College Composition and Communication (CCCC) committee chair

Beth Hewitt, wrote a cautiously optimistic review of the AES system, WriteLab, and reported that despite legitimate concerns about machine graders designed to replace human readers, "WriteLab's current configuration and stated goals should not be ethically troublesome for writing center educators" (Hewitt, 2016). In 2015, Turnitin ran a pilot study on their Revision Assistant system with 18 middle and high schools in the United States (Turnitin, 2015). 94% of students revised their work at least once, compared to an earlier study using Criterion system, in which only 29% of students revised their work. The study also revealed the average word counts of students' work gradually increased with each revision. In addition, students' grades increased after rewriting their work. As a result, follow-up interviews revealed positive attitudes from both students and teachers regarding the use of Revision Assistant.

In a classroom-based study of English Language learners at a university in Taiwan, Chen and Cheng (2008) analyzed perceptions of MY Access! used by students and instructors in three different classes. They found that instructors' attitudes to the software and the way it was used greatly impacted students' perceptions. When teachers had a more positive attitude towards the system, students did as well. Furthermore, when scores and feedback from MY Access! were combined with teacher and peer feedback, and when the AES system was used for formative rather than summative assessment, students' attitudes were more positive.

In a qualitative study investigating the practices and perspectives of five university teachers implementing the AES system, Criterion, in ESL writing courses, researchers found Criterion produces positive results (Link et al., 2014). However, the researchers were careful to point out that all the teachers involved in the study were considered proficient users of the technology and thus felt comfortable integrating it as a classroom tool. This study indicated

successful AES implementation relies on teachers' willingness to ask questions, explore the technology, and be flexible in adapting its use in ways that best fit the context. The authors of the study conclude by calling for future research to focus on the effect of training teachers to use and integrate AES tools in the writing curriculum.

Overall, perception studies suggest careful use of AES systems to supplement writing instruction may lead to positive attitudes from both teachers and students. When teachers understand the systems are designed to augment, not replace, their feedback, they are more receptive to trying new systems. When students use AES as a formative assessment tool that offers immediate feedback and helps them improve subsequent essay drafts, they tend to produce longer and more frequent revisions, suggesting higher levels of motivation. Effective integration of AES tools will depend on teachers' perceptions of and familiarity with the technology. If teachers do not explore the tools and understand the technology, they will not know how to implement it effectively.

Studies Using LightSide as an AES System

In the field of ESL, widely used AES platforms include Write&Improve from Cambridge English, Criterion from ETS, WriteToLearn from Pearson, and My Access! from Vantage Learning (Hockley, 2018). These commercial platforms are costly and out of reach for many ESL programs. Moreover, these platforms were not originally designed to analyze learner English (Ranalli, 2017). Machine learning offers a solution in that it can "learn" the patterns unique to ESL writing and scoring by extracting features from human-graded training sets. In other words, through machine learning, ESL teachers can build their own customized, homegrown AES system using their students' graded essays as input. Access to machine learning, however, has

been exclusive to developers and technical experts due to the complex nature of computer programming and statistical models (Mayfield & Rosé, 2014). Researchers at Carnegie Mellon University developed LightSide as an open-source, machine learning platform developed with non-expert users in mind.

With LightSide, developers Mayfield and Rosé (2014) claim to break down what they refer to as the "black box" of machine learning into smaller component parts, simplifying the workflow, and demystifying the process of text analysis. Because it is open-source, users can explore various models and customize feature extraction according to their specific tasks, and experienced programmers build new plug-ins and test new feature representations without constructing a platform from scratch. The developers emphasize LightSide is a flexible tool capable of performing a variety of text mining tasks, including sentiment analysis, data annotation, text classification, and automated essay scoring. This section of the literature review presents studies that used LightSide for the specific task of essay scoring.

Holistic Essay Assessment

LightSide gained attention when the developers participated in the 2012 competition, organized by Kaggle and sponsored by the Hewlett Foundation, with the purpose of comparing state of the art AES systems to inform policy and decision making for stakeholders in essay assessment (Shermis & Hamner, 2013). The only open-source AES platform, LightSide competed against eight commercial vendors who, collectively, represented over 95 percent of the market (Mayfield & Rosé, 2014). The competition challenged participants to build a model trained on data for eight different essay prompts. The competition organizers provided a dataset ranging from 900 to 1800 student essays for each of the eight prompts, all of which had

been scored by at least two human raters. Agreement findings were reported using Pearson *r* correlation, with human agreement ranging from 0.61 *r* to 0.85 *r* (Shermis & Hamner, 2013). LightSide's performance ranged from .64 to .81, with an average of 0.75 *r* across datasets, exceeding inter-rater agreement between humans on five of the eight prompts (Mayfield & Rosé, 2014). On all prompts, LightSide performed as well as, and in some cases even better, than most of its commercial competitors (Shermis & Hamner, 2013). Findings from this study suggest LightSide is a promising tool for holistic essay assessment.

Trait-Based Essay Assessment

Some studies have investigated LightSide's potential for scoring whether students' essays contain a specific concept covered in the curriculum. One such study examined whether LightSide is capable of evaluating college biology students' written explanations of evolutionary change (Ha et al., 2011). The study used a dataset of 2556 short essays written by students at Ohio State University and Michigan State University. Two human raters scored the essays for the presence or absence of five key concepts of evolution: variation, heredity, limited resources, competition, and differential survival. The study found that LightSide was highly effective at scoring the accuracy and complexity of students' explanations of evolutionary change, although the researchers found a few limitations that confused the scoring models. Limitations included misspelled words, nonadjacent key terms, uncommon concept frequencies, and the diversity of expressions students used to explain some concepts. Overall, the results indicated LightSide is an effective tool for trait-based assessment.

Collaborative Learning Tool

In a study on collaborative learning discussions, researchers analyzed chatroom conversations from an undergraduate college course (Howley et al., 2012). Chatroom transcripts were uploaded and annotated in LightSide, with the goal of understanding the effect of group composition on student behavior and self-efficacy. The results of the study showed that LightSide is not particularly effective in annotating the short utterances consisting only choppy phrases and incomplete sentences often used in chat. A new model had to be designed with new features such as length between turns in discourse, and vocabulary similarity from one response to the next. With extensive human-labeled examples to train a new model, LightSide can be useful in annotating online conversations.

Summary of LightSide Research

LightSide is a useful tool that could potentially allow teachers to build their own AES system trained entirely on their students' essays as input. Though this is a powerful tool, it does require training. To date, a limited number of studies have been published on the use and effectiveness of LightSide as a classroom tool. To the author's knowledge, no studies have been published on using LightSide as a tool to evaluate learners of English.

Summary of Research

Automated scoring systems have advanced considerably. Machine learning techniques make it possible to train new scoring models to mimic human scoring efforts. Several studies have pointed to the utility of machine learning in educational assessments as a tool for quickly and accurately scoring written text (Zhai et al., 2020). LightSide offers a user-friendly, open-

source platform that can be used for both holistic and trait-specific scoring (Mayfield & Rosé, 2014). Several studies over the past decade have indicated LightSide produces highly accurate results, consistent with human scores. However, nearly all published research on LightSide looks at writing samples from native or near-native English speakers, with very little attention given to learner English.

While there is a growing body of research pertaining to AES systems used in the context of ESL/EFL education, several questions remain, providing potential avenues for future research. One question regarding current AES systems is whether the grammar, usage, and vocabulary features used to score essays are useful features for evaluating the language of non-native speakers. Another question is whether AES developers have enough reliable learner corpora to train and test AES systems. The training, calibration, and expertise level of human raters raises important questions about the reliability of testing data. Despite findings that AES systems can be as reliable and valid as human scores, usage continues to generate controversy in the teaching community, mainly because writing teachers guide students in composing texts for an audience, and a computer is not a "real" audience capable of reading students' work (Herrignton & Moran, 2001; Ansen, 2006, Weigle, 2010). To build trust among ESL educators, AES researchers and developers should take measures to construct learner corpora for training and testing systems, ensure reliable and valid scoring methods that take linguistic differences into account, and consider curriculum alignment for use in language education.

CHAPTER 3

METHODOLOGY

This study aimed to understand ESL teachers' initial attitudes towards using automated essay scoring and to describe the effect of training on teachers' attitudes. The training consisted of a two-hour workshop on how to use LightSide, a freely available open-source software, as an automated scoring tool. The study was guided by the following central questions: How do ESL teachers feel about using AES to enhance instruction? What challenges did teachers face in learning to use LightSide? How did the completion of a brief workshop change teachers' perceptions? How do teachers intend to use LightSide after completing the training?

Research Design

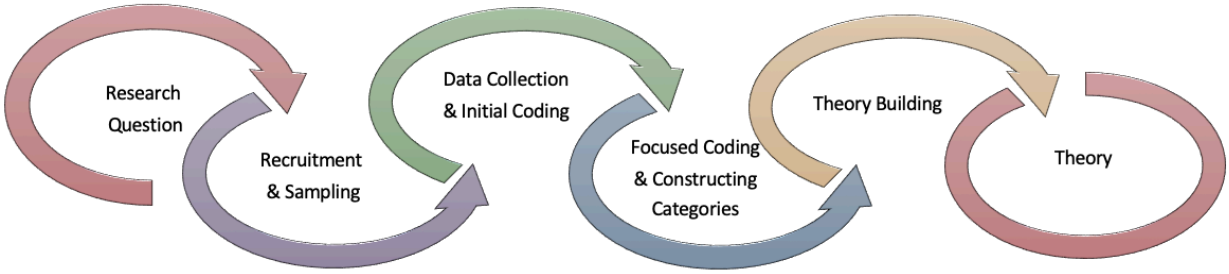
This study took a qualitative approach to allow the researcher to explore and understand the meaning of the problem as it is perceived by both the individual teachers and the group at large. Creswell (2014) defines qualitative research as a method of inquiry that values an inductive approach, relies on emergent questions and processes, builds from specific to broad themes, and emphasizes personal meaning, to accurately depict the complexity of a situation. This type of research involves asking questions and crafting procedures to extract information and opinions from a designated group of people as it relates to a specific problem (Creswell & Creswell, 2018). The problem in this case was ESL teachers' resistance in implementing AES systems for instructional use. The research methods were designed to extract teachers' firsthand perceptions, knowledge, and experience in using AES.

The study adopted the grounded theory framework in an inquiry-based approach to

research based on participants' understanding. First developed by Glaser and Strauss (1967), the defining features of grounded theory include simultaneous involvement in data collection and analysis, constructing analytic codes and categories from data, constantly making comparisons throughout each stage of the analysis, writing memos to specify and elaborate categories, and developing theory during each step of data collection and analysis. Grounded theory uses methodical yet adaptable guidelines for gathering and examining qualitative data to develop theories grounded in the data (Charmaz, 2014). This method begins with inductive data, uses comparison approaches, and employs iterative strategies for switching back and forth between data and analysis. Figure 3.1 provides a graphic representation of the cyclical process used in grounded theory to compare, organize, gather, and analyze data (Corbin & Strauss, 2015).

Figure 3.1

Grounded Theory Model



Participants

Data collection in grounded theory research typically involves theoretical sampling techniques in which data are gathered in iterative waves to recruit participants who can provide insight to the emerging theory. Both purposeful and convenient sampling methods

were used to select participants for this study. Purposeful sampling allowed the researcher to select subjects who are experienced in teaching writing to ESL students in a secondary or higher education setting. Convenient sampling was employed because the researcher, also an ESL teacher, was able to identify suitable participants through professional networking such as conferences and both past and present employment in ESL programs.

The participants included eighteen teachers – eleven women and seven men – all of whom have taught ESL in a higher education setting for at least five years. Fifteen of the participants have taught for more than ten years. Thirteen of the participants are currently teach in Intensive English Programs at community colleges, four teach in universities, and one teaches at an adult literacy center. All but one of the participants rated themselves as proficient users of technology.

Procedure

The study was approved by the Institution Review Board of the University of North Texas (IRB-22-375). Participants were emailed information about the study, including the purpose of the study, a description of the procedures, information about the benefits of participation, contact information of the researcher, a statement that participation is voluntary and may be withdrawn at any time, and an informed consent form.

Participants who consented received a pre-workshop survey to complete a week before the workshop. Along with the survey, participants received step-by-step instructions for downloading and installing the software to be used for the workshop. Participants attended via Zoom a two-hour workshop, in which they were introduced to the open-source platform, LightSide, and shown how to use LightSide for the purpose of AES. Immediately following the

workshop, participants were invited to attend a focus group interview to provide initial feedback about their perceptions of using LightSide. Delayed semi-structured interviews took place a week or two after the workshop to assess general changes in teachers' perspectives after they had explored the software and gained more experience on their own.

Pre-Workshop Survey

Surveys were disseminated through Qualtrics to collect participants' attitudes towards using technology to enhance language instruction, their initial perceptions of machine learning and AES systems, and their willingness to implement AES for classroom use. A copy of the full survey is included in Appendix A. The following questions were included:

- How do you integrate technology in your ESL writing classes?
- Do you have any experience with machine learning, artificial intelligence, and learning analytics? If yes, explain more.
- Are machine learning, artificial intelligence, and learning analytics useful for teachers?
- Do you have any experience using automated essay scoring systems in the past? If yes, explain more.
- Is automated essay scoring useful for teachers?
- What are the pros and cons of using an automated essay scoring system?
- What are your expectations in terms of potential for automated essay scoring and the ways it will impact your teaching?

Teacher's Workshop

I designed the workshop, "Exploring LightSide: A Workshop for ESL Instructors," utilizing the open-source text mining application, LightSide, along with a large dataset of authentic essays composed by ESL students for the TOEFL exam. A quick-reference guide was created for

workshop participants (see Appendix B). Conducted through the videoconferencing platform, Zoom, the workshop spanned roughly two hours. It began with an overview of the fundamental concepts of machine learning, elucidating the steps involved in the machine learning workflow. The session then delved into the specifics of AES, its underlying workflow, and the mechanics of how it functions within the context of language education. Participants were then guided through potential classroom applications of AES, as well as the advantages and disadvantages of employing AES, as detailed in Table 3.1. This approach allowed participants to gain a well-rounded perspective on the topic, equipping them with the necessary knowledge to make informed decisions about the use of AES in their classrooms. The rest of the workshop guided teachers to open LightSide on their own computers to learn how to use the software, as illustrated in Figure 3.2.

Table 3.1

Pros and Cons of AES in Language Education

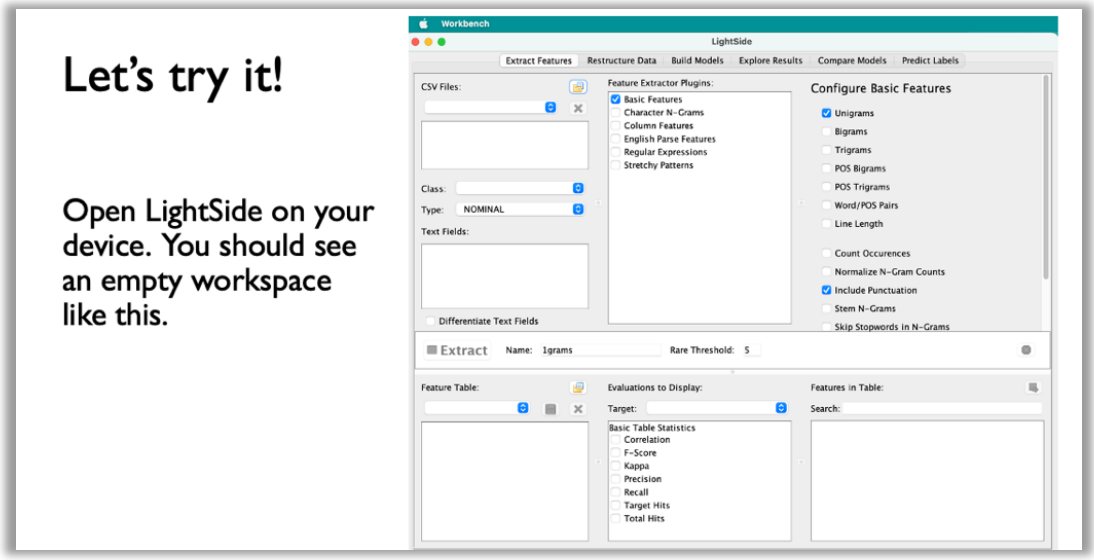
Pros	Cons
• Immediate scores; quick identification of at-risk students • Frees up teachers' time to focus on tasks other than scoring • High validity and reliability • Eliminates human error factors (fatigue, distraction, bias, subjectivity, psychology, etc.)	• Student anxiety • Lower motivation • Commercial systems not designed for learner (non-native) writing • High costs for specialized systems • Essay collection/corpus creation are time-consuming and difficult • Vulnerable to cheating

Participants downloaded and installed the application on their devices before attending the workshop. This interactive component of the workshop lasted approximately 90 minutes,

during which time participants learned how to load new data, extract features, train models, interpret results, improve models, and score new, ungraded essays. Participants were given opportunities to ask questions and comment on their experiences during the workshop.

Figure 3.2

Interactive LightSide Training



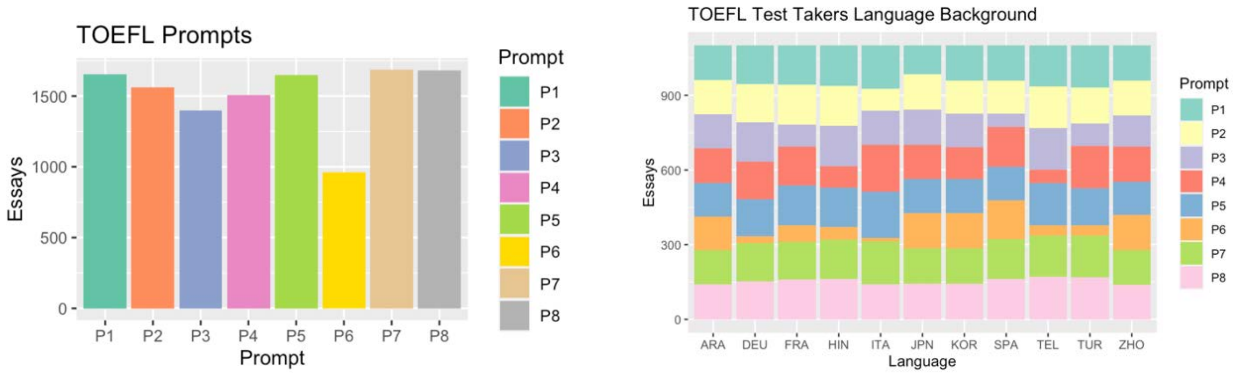
Workshop Materials: TOEFL11

The dataset used for "Exploring LightSide: A Workshop for ESL Instructors" comes from TOEFL11, a publicly available corpus of essay samples written by non-native English speakers (Blanchard et al., 2013). The dataset consists of 12,000 authentic essays written by test takers for the TOEFL, a high-stakes college-entrance test designed to assess non-native speakers' academic English proficiency. The test is administered by Educational Testing Services (ETS), a private nonprofit educational testing and assessment organization. ETS released the TOEFL11 dataset to promote research in the fields of computational and corpus linguistics (Blanchard et al., 2013).

The TOEFL11 test takers took the TOEFL on computers at secure ETS testing centers around the world from 2006-2007. The TOEFL is used internationally to assess college-level academic English proficiency to help institutions of higher education make admissions decisions. The test consists of reading, writing, listening, and speaking sections and typically takes about four hours to complete. As of 2022, the cost of the TOEFL ranges from $180-$325, depending on location. Because the test involves a significant time and financial commitment, it may be assumed that most test-takers have prepared for the exam and are seriously seeking admission to colleges and universities in which English is the primary language of instruction. The test takers came from eleven language backgrounds: Arabic (ARA), German (DEU), French (FRA), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Chinese (ZHO), with an even distribution of 1,100 per language group.

Figure 3.3

Distribution of Languages and Prompts in TOEFL11



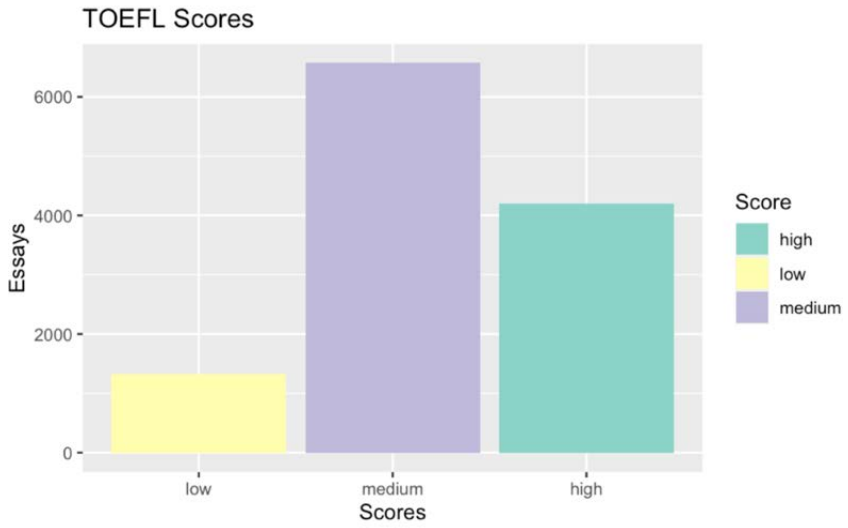
The writing task is scored on a 5-point scale, according to a rubric provided on the ETS website. Each essay in this dataset was scored twice, by two human raters. (After the release of the TOEFL11 dataset, ETS changed its scoring system to be graded once by a human rater and once by e-rater, a propriety AES system designed by ETS.) The two raters' scores were averaged

if the scores differed by more than one point. An essay that assigned a score of 3 from two human raters received a final score of 3. An essay assigned a 3 from one rater and a 4 from a second rater received a final score of 3.5. If the two raters' scores differed by more than one point, a third human rater would assign a score. If the three scores were adjacent, the average was used. For example, if the three scores consisted of 3, 4, and 5, the final score would be 4. If one score was an outlier, the two adjacent scores would be averaged. For example, if the three scores consisted of 2, 3, and 5, the final score would be 2.5 since the high score was an outlier. If the three scores were 1, 3, and 5, a fourth rater would provide an adjudicated score.

The average of the two human raters were collapsed from the original 5-point scale to a 3-point scale consisting of low, medium, and high labels, with "low" applied to essays scored from 1 to 2, "medium" for essays scored from 2.5 to 3.5, and "high" for essays scored from 4 to 5. The scores represented in the dataset are not evenly distributed, as shown in Figure 3.4, with 6,568 medium scores, 4,202 high scores, and only 1,330 low scores represented.

Figure 3.4

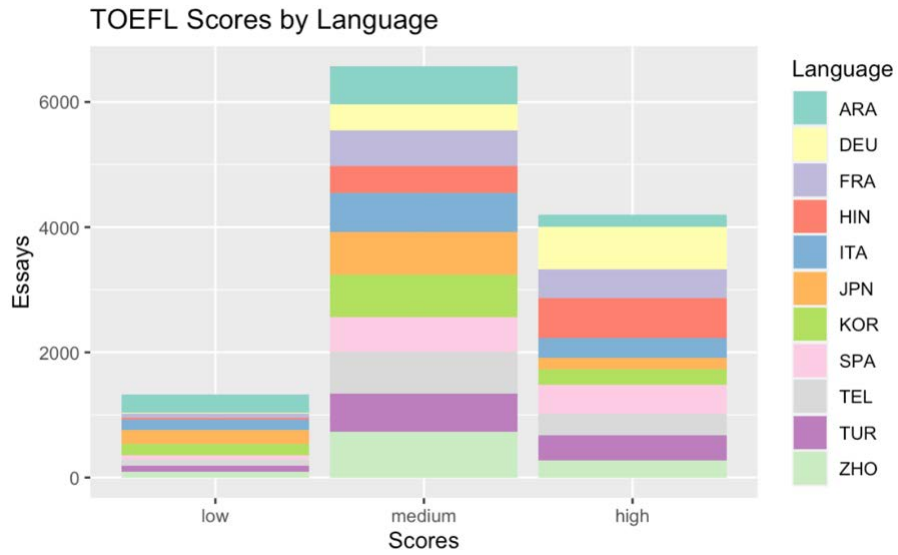
Distribution of Scores in TOEFL11



The test takers' language background is somewhat evenly distributed across the medium scored essays. Languages are represented more disproportionately across the high and low scores, as shown in Figure 3.5.

Figure 3.5

Distribution of Scores in TOEFL11



Workshop Tool: LightSide

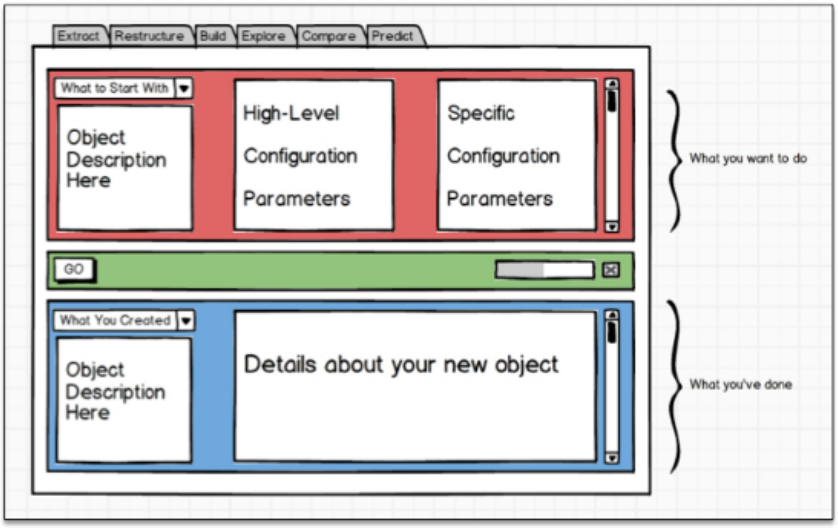
From the wide variety of AES tools currently available, I selected LightSide for this study for several reasons. First, LightSide is not a pre-built AES system but a text-mining platform that enables users to build their own models. This aspect is important because pre-built models are generally not tailored for assessing learner English. LightSide allows ESL teachers to train models using their own students' essays and extract relevant features found in their students' writing. In contrast to proprietary and expensive pre-built AES systems, LightSide is open-source, making it freely available and customizable. Given that machine learning typically requires an in-depth understanding of mathematics and computer science, creating a steep learning curve for novice users, perhaps the most important reason for selecting LightSide may

be its user-friendliness. The training manual states, “We’ve built a tool that lets you hit the ground running with your data, putting as much of the research workflow for machine learning as possible into an easy, point-and-click interface” (Mayfield et al., 2014, p. 1).

LightSide’s workspace is organized into six tabs: Extract, Restructure, Build, Explore, Compare, and Predict. The most straightforward workflow involves only the Extract and Build tabs. In the Extract tab, documents are loaded, and features are extracted to create a feature table. The Build tab enables the selection of algorithms to create a model that classifies results to replicate human labels. Figure 3.6 illustrates this basic workflow. The Restructure tab provides tools for manually adjusting feature tables. The Explore tab contains analysis tools for better understanding the models. The Compare tab allows users to look at two different trained models side by side. Finally, the Predict tab allows users to input new, ungraded writing samples to be scored.

Figure 3.6

LightSide Workflow



Source: Mayfield et al. (2014).

Feature Extraction

The "Basic Features" settings on LightSide allow users to configure standard text features to be extracted from the data. N-grams represent words. The most basic feature is the unigram, which simply checks for the presence or absence of a single word. Bigrams and trigrams consist of two or three adjacent words, respectively. Bigrams and trigrams are important for catching common phrases and collocations in texts. POS (or part of speech) n-grams represent the syntactic function of words such as verb tense, singular and plural nouns, pronouns, prepositions, and so on. LightSide uses the Stanford POS tagger (Toutanova et al., 2003) developed by computational linguists at Stanford University, to annotate part of speech. An example of a POS bigram might look like this: *PRP_VBP,* where *PRP* represents a personal pronoun, such as *I, you, he, she,* or *they,* and *VBP* represents a non-third person singular present tense verb. For example, the phrase *they look* would be annotated as *PRP_VBP.* A POS trigram contains three POS n-grams, such as *PRP_VBP_JJ,* where *JJ* represents an adjective. *They look happy* would be annotated in this way.

The "Line Length" feature captures exactly what it sounds like, a single feature representing the number of words in a text. Whereas n-grams are represented with a "true" or "false" value indicating the presence or absence of words and phrases, line length is represented with a numeric value. The "Count Occurrences" feature can be used to assign a numeric value for n-grams, representing how many times a word appears in a text. The feature "Normalize N-Gram Counts" normalizes the numeric values assigned to n-grams by indicating the proportion of the document each word represents. These features might be useful if teachers are looking for coherence, in which case a high count of key words might lead to a

better essay score, or if they are looking for vocabulary variety, in which case a low count of each n-gram might be preferrable. In other words, the "Count Occurrences" and "Normalize N-Gram Counts" features might be useful for very specific tasks, but probably not for the holistic scoring of large-scale placement exams.

The punctuation feature checks for commas, periods, question marks, quotation marks, apostrophes, and so on. When testing data to be used for the workshop for this study, checking or unchecking the punctuation feature did not seem to have a big impact on the scoring models. However, because mechanics is an important component of any writing class, teachers will most likely want to see this information.

Stemming n-grams reduces words to their simplest form, similar to, but less extreme than, lemmatization, a process which groups related words to a single form, or lemma. Stemming basically removes morphemes such as prefixes and suffixes to reduce words to their root forms. For example, the verbs *take, takes, took, taken, and taking* would all be represented simply as *take*. Stemming might be useful for trait-based tasks, such as summary writing or focused response questions, but important grammatical features of words will be lost.

Stopwords consist of the most commonly used words that do not have any meaning on their own but serve a grammatical function. Examples include *a/an, the, and, but, is, are, to,* and *for.* LightSide provides three feature extraction options for dealing with stopwords. "Skip Stopwords in N-Grams" completely passes over stopwords. For example, "My sister is a student" would be represented only as "sister_student". This option might be useful if the task is more concerned with content than style and mechanics. "Ignore All-stopword N-Grams"

removes unigram stopwords but keeps stopwords in bigrams and trigrams if the bigram or trigram also contains non-stopwords. “Contains Non-Stopwords” provides a binary true/false value based on whether the text contains a single non-stopword (true) or not (false). Since it is unlikely an essay could contain zero non-stopwords, this feature is not useful for essay scoring.

An additional option within the Extract tab is the ability to adjust the rare threshold, which controls the number of times a feature must appear in a text for it to be included in the algorithm. By default, LightSide sets the rare threshold to five, which means a feature must appear in a dataset at least five times. This feature is useful for extremely large datasets. A feature that shows up only five times in a dataset of thousands of essays, for example, might not be important. In this case, the rare threshold could be increased.

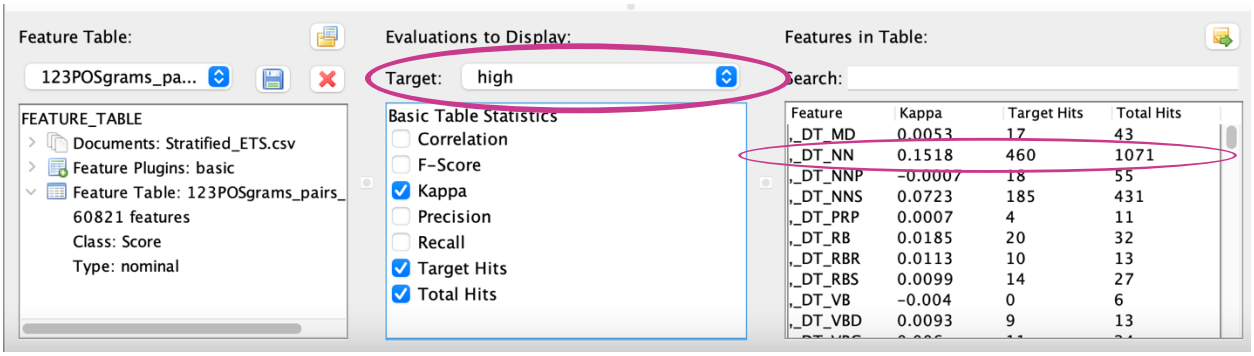
Several other feature extractor plugins are available, including “Regular Expressions”, an option that allows users to specify text patterns, “Stretchy Patterns”, a tool for extracting words that are close together but not adjacent, “Character N-Grams”, which extracts strings of characters rather than entire words, and “Parse Features”, which extracts production rules and dependency relations to identify how words in sentences are grammatically related. These advanced extraction tools are best suited for experienced users and will not be presented in the teachers’ workshop.

After selecting the features to be extracted, users can view a feature table, as illustrated in Figure 3.7. The box on the left provides general information about the dataset and the total number of features that were extracted from the dataset. The middle box allows users to select different types of statistics. For the purposes of the teachers’ training workshop, “Total Hits”, and “Target Hits” will be selected. Total hits represents the number of essays in the entire

dataset that contains each feature. Target hits shows the number of essays in a given category (score). The dataset used in the training classifies students' essays into three categories, or scores: low, medium, and high, according to the level of English proficiency test-takers demonstrated on the TOEFL essay. Users can select one of the three scores from the "Target" dropdown menu above the box for statistics selection. The right box shows the feature table, with a complete list of features extracted from the essays in that target classification, along with the selected statistics. Figure 3.7 shows the feature DT_NN, which represents a determiner such as "this" or "the" followed by a singular or mass noun such as "coffee" or "dissertation", occurs in 1,071 of the 3,000 essays (total hits) and in 460 of the 1,000 essays classified as *high* (target hits).

Figure 3.7

LightSide Feature Table



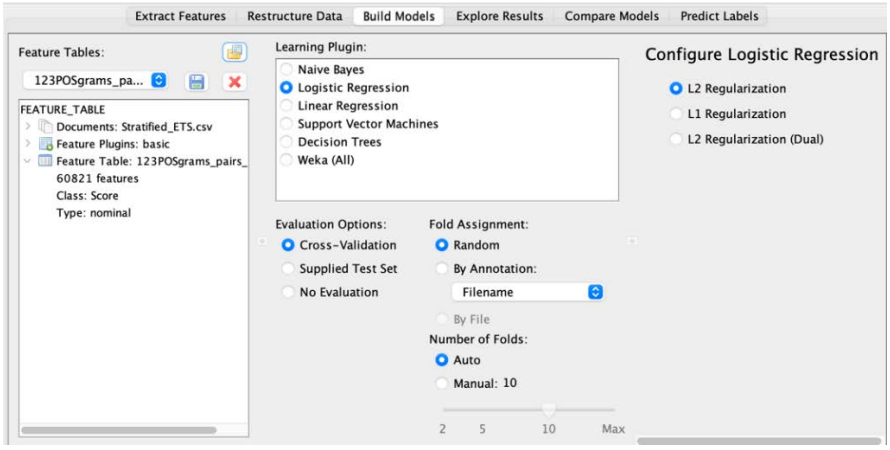
Machine Learning Algorithms

After building a feature table, the next step is to train a model using a machine learning algorithm. This may be the most intimidating step for ESL teachers, especially those with little or no background in statistics. Fortunately, the LightSide developers have made this step very simple, with a clickable list of the most commonly used algorithms, as illustrated in Figure 3.9.

Even those who have no knowledge of statistics will be able to click an algorithm, build the model, see the accuracy score, and then try another algorithm for comparison. For the TOEFL dataset, the most effective algorithm is Naïve Bayes, a simple but highly effective classifier that predicts labels on the basis of probability. The second most effective algorithm for this dataset is logistic regression, a linear classifier. The Naïve Bayes and logistic regression algorithms are the only two models that will be presented in the teachers' workshop. Whether one model is more effective than the other depends on the dataset. It will be useful for teachers to learn how to explore and compare results from the two classifiers.

Figure 3.8

LightSide Algorithms



Building an AES system typically requires a separate dataset used to test the validity of the model. To do this in LightSide, users must select the "Supplied Test Set" in the area beneath the algorithm options. This validation technique will not be used in the workshop as it would require teachers to manage two separate datasets. LightSide offers a less cumbersome approach, cross validation, which allows users to enter one single dataset for building *and* validating the model. Cross validation randomly splits the data into several sets, known as

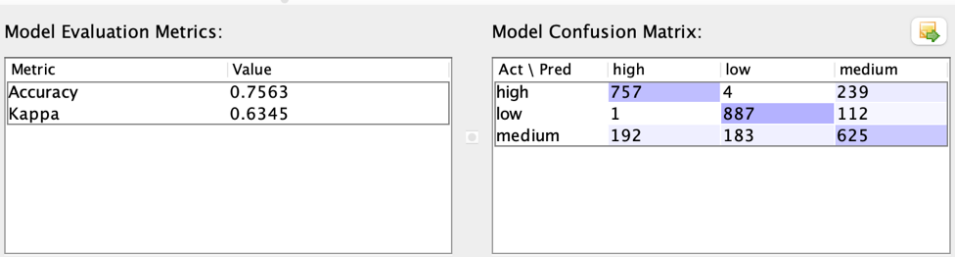
"folds." The default setting is ten folds, which means the dataset is split into tenths. The dataset is processed ten times, with nine of the ten sets of essays used to train the model and the tenth used for validation.

Interpreting Results

After building and training a predictive model, users will see the model's accuracy and Kappa, as shown in Figure 3.10. The accuracy represents how many essays it scored correctly. In the example below, almost 76% percent of 3,000 essays were accurately scored. Kappa represents how well the model performed above chance. In addition to these broad metrics, users can see how well the model predicts each score by examining the model confusion matrix. In Figure 3.9, the rows *high, low,* and *medium* indicate the actual score, while the columns represent the predicted scores. In this example, the model correctly predicted 757 high scores, 887 low scores, and 625 medium scores.

Figure 3.9

LightSide Evaluation Metrics and Confusion Matrix



Scoring New Essays

After extracting features and training a model, teachers now have their own homegrown AES system built from their own dataset and programming choices. At this point, teachers can use their model to score new essays. This is done through LightSide's "Predict

Labels" interface. During the workshop, teachers were supplied with a second dataset of 1000 random TOEFL essays that were not part of the 3000-essay set used to build and train their model. Teachers were guided to load the new file, process the data, and view the predicted scores.

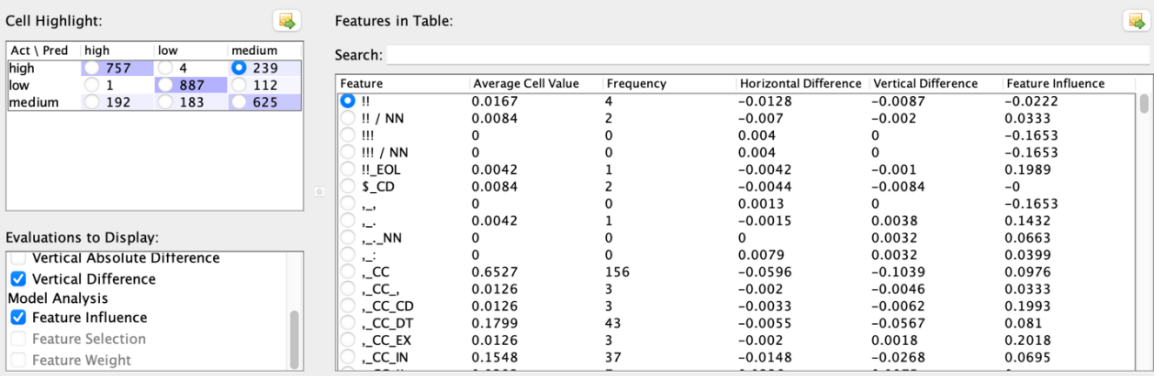
Analyzing Errors

The most basic workflow in LightSide consists of using only three interfaces: "Extract Features", "Build Models", and "Predict Labels". Teachers will use "Extract Features" to load new data, select features, and build a feature table. Next, teachers used "Build Models" to choose an algorithm and validation method to train a predictive model. Once the model is built, new essays can be loaded and scored by the model, using the tab "Predict Labels". While this by itself is useful, teachers need more information, especially considering the model is not 100% accurate. If a model is 76% accurate, teachers should ask what went wrong with the 24% that missed the mark.

LightSide provides the interface "Exploring Results" for users to gain a deeper understanding of the data. Here, teachers can view the distribution and frequency of specific features according to their actual and predicted scores, as shown in Figure 3.10. It is useful to view the largest group of essays that were scored incorrectly. In this example, 239 of the 1000 high-scored essays were incorrectly labeled as medium. Teachers can look at the feature table for that set of essays to see what features define those essays. In this new feature table, teachers can view the frequency of each feature, or how many essays in a given set contain that feature. The average value column shows the percentage of essays in the set that contain feature.

Figure 3.10

Exploring Results in LightSide



Some advanced tools in this interface allow teachers to understand which features are the most prominent in "confusing" essays, or those that the algorithm could not correctly predict. Horizontal Difference calculates the difference in the average value of a feature that occurs in a confusing essay versus the average value of that same feature in correct predictions. This allows teachers to understand which features are the most different between essays that LightSide correctly labeled and those where the machine learning made a mistake. Vertical Difference provides a similar calculation, but it focuses on the similarities between the different cells rather than the differences. Influence measures the influence of a feature by calculating how different the classification would be if that feature were added or removed. While Horizontal Difference, Vertical Difference, and Influence offer insightful information for understanding and improving the model, they are too complex for beginners in machine learning, the target audience of this workshop. Thus, these tools will be mentioned only briefly for teachers who are interested in learning to refine and improve their AES model, but they will not be presented in depth during the workshop.

The bottom half of the Explore Results interface provides tools that allow teachers to

easily identify which specific essays the model predicted correctly or incorrectly. The plugin "Label Distributions", shown in Figure 3.11, creates a row for each essay, along with columns representing the actual scores and predicted scores. This is useful for quick identification of any outliers, or essays that were incorrectly classified by LightSide.

Figure 3.11

Label Distributions

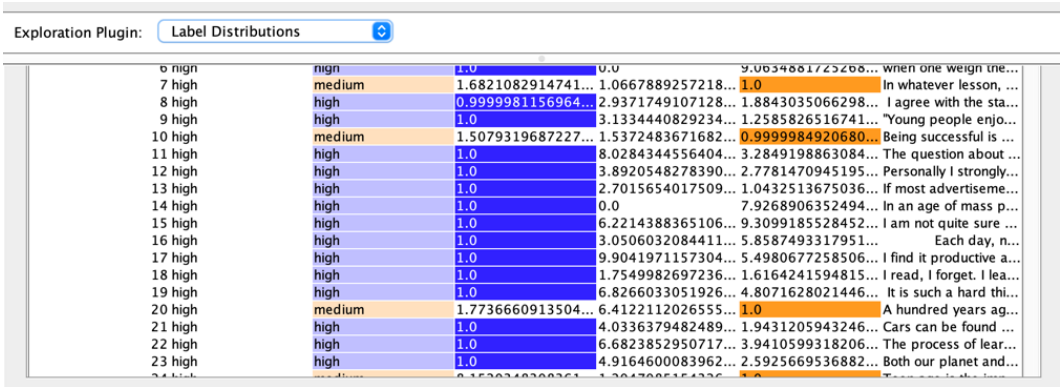
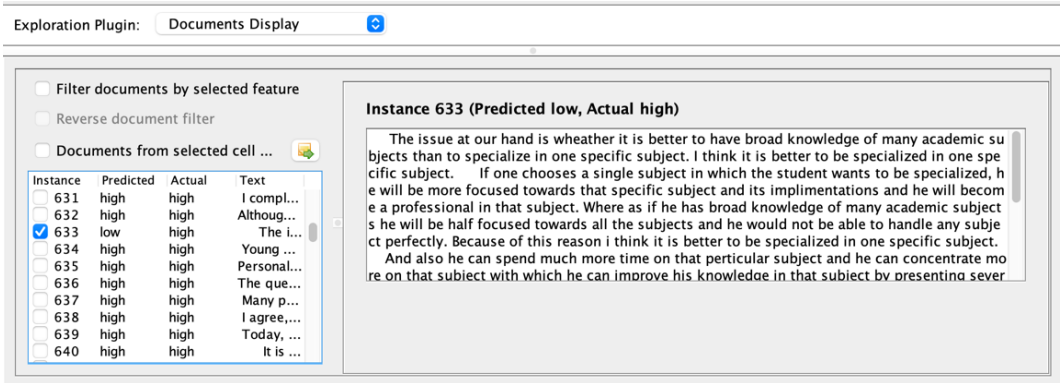


Figure 3.12

Analyzing Essays



After identifying which essays confused the model, teachers can use the "Documents Display" plugin to view an entire essay. Figure 3.12, for example, shows an essay that was given a *high* score by a human grader but classified as *low* by LightSide. During the workshop,

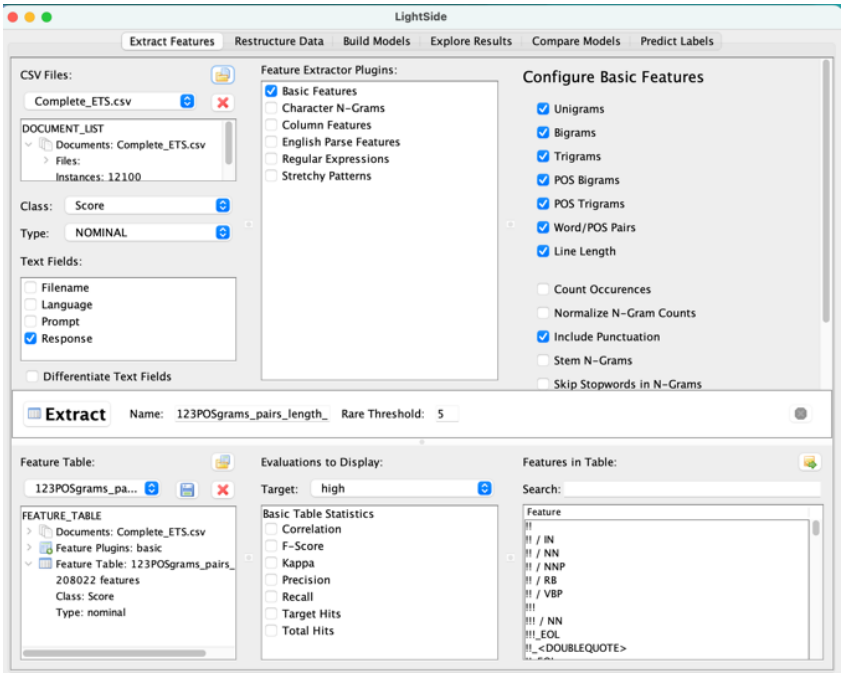
teachers examined four essays that were misclassified. Teachers were then asked their opinion about which grader was “off”, the human or the machine.

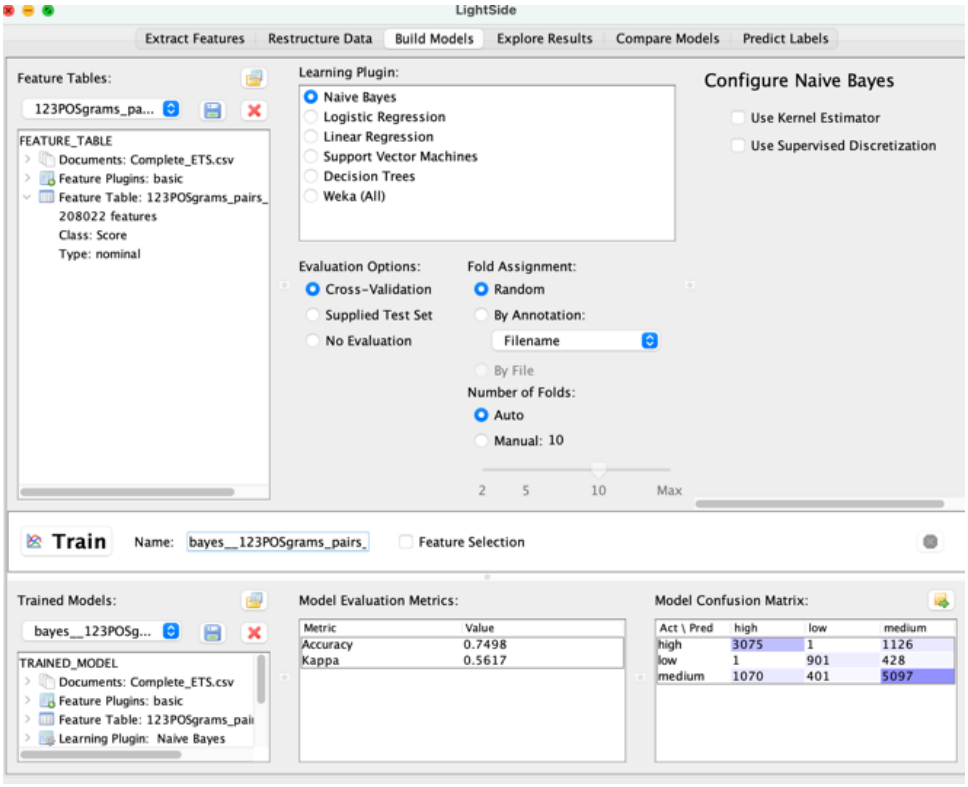
LightSide Results Assessing TOEFL11

Beginning with the entire TOEFL11 dataset of 12,100 essays and employing a few of LightSide’s basic extraction features – unigrams, bigrams, trigams, POS bigrams, POS trigrams, POS pairs, line length, and punctuation – yielded over 200,000 features and required substantial memory for processing. Using a Naïve Bayes method, the model achieved approximately 75% accuracy with a kappa of 0.56. Modifying the features and increasing the rare threshold to extract fewer features did not significantly impact the model. As these features and the Naïve Bayes algorithm produced the most optimal results, illustrated in Figure 3.13, workshop participants were instructed to choose these options.

Figure 3.13

LightSide Results on 12,000 TOEFL11 Essays





To make the data more manageable, the sample size was reduced to 3,000. Because the goal is to determine how well LightSide can predict scores for ESL essays, it makes more sense for the dataset to be stratified by score rather than language background. Therefore, 1,000 essays from each score, *high, medium*, and *low,* were randomly selected. The prompts in the new essay sample have a similar distribution, illustrated in Figure 3.14, to that of the original dataset.

The languages are not evenly distributed in the new sample (see Figure 3.15). This is unavoidable because the languages were not evenly distributed across scores in the original sample. Since the original sample contained only 1,330 essays with a *low* score, most of those essays were included in the new sample representing 1,000 essays for each score. The original sample of *low* essays included a disproportionately large number of test takers from an Arabic

or Japanese language background. The new sample follows a similar pattern. However, the new sample still contains over 200 essays from each of the eleven languages, providing a good representation of all the language groups.

Figure 3.14

Distribution of Prompts in 3,000-Essay Sample

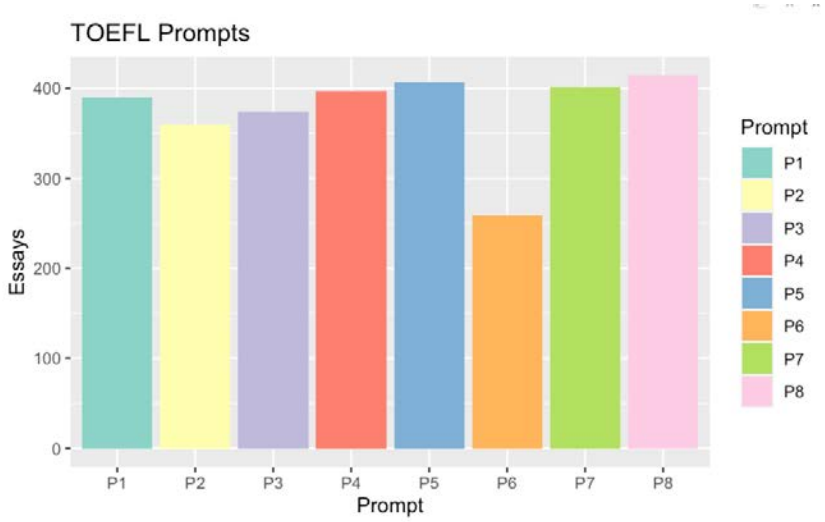
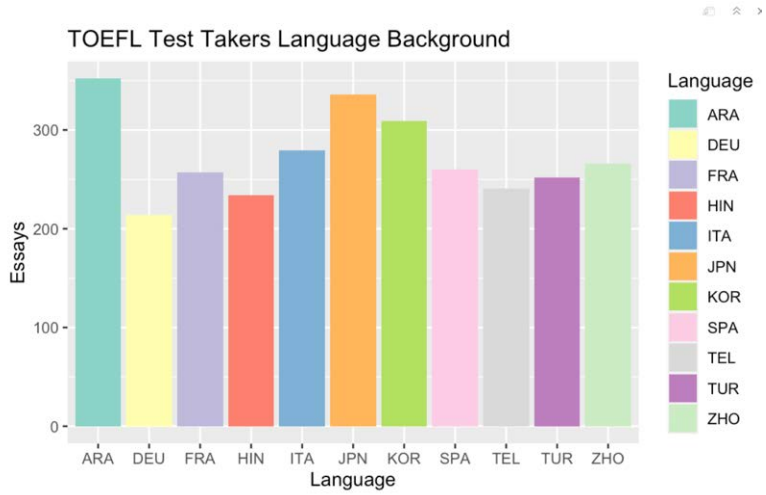


Figure 3.15

Distribution of Languages in 3,000-Essay Sample

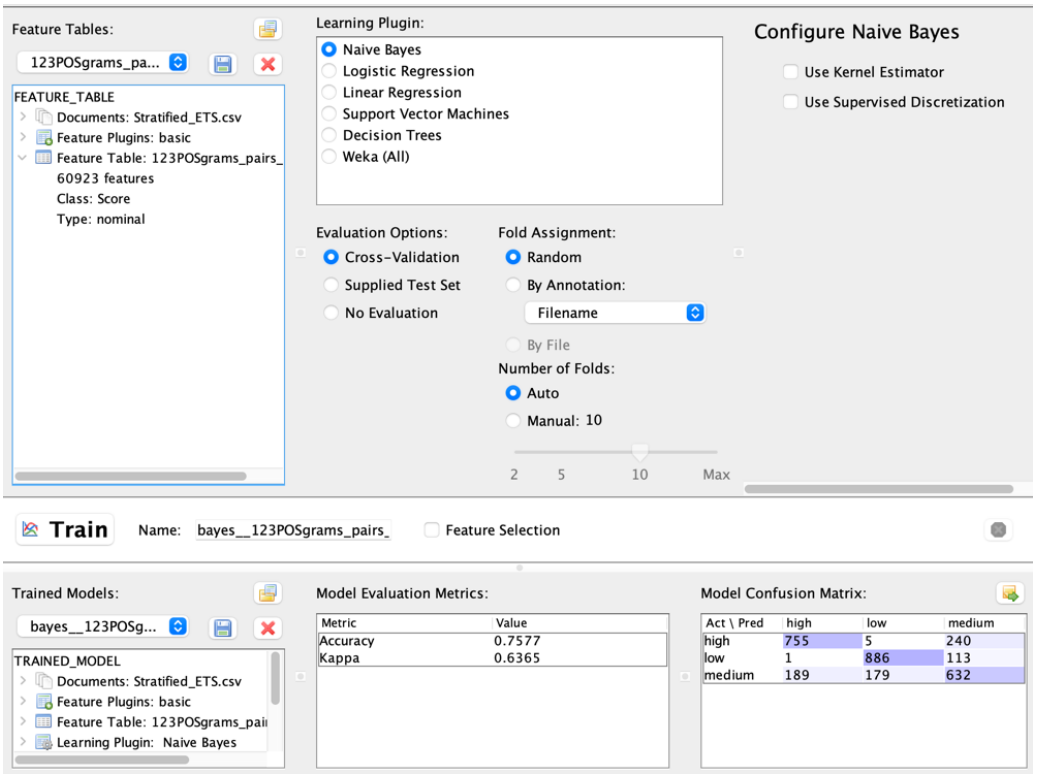


Language <chr>	n <int>
ARA	352
DEU	214
FRA	257
HIN	234
ITA	279
JPN	336
KOR	309
SPA	260
TEL	241
TUR	252

By choosing the same basic LightSide feature extraction employed previously for the complete (12,000-essay) dataset, the new 3,000-essay dataset yielded just over 60,800 features, rendering it much more manageable and efficient to work with. Interestingly, when using the same Naïve Bayes model as before, the accuracy and reliability statistics were considerably higher for the smaller dataset, with 76% accuracy and a kappa of 0.64. These outcomes support the decision to use the smaller dataset for the teachers' workshop, as it delivers more reliable results while also being less cumbersome to handle.

Figure 3.16

LightSide Results for Smaller Sample



For the workshop, the 3,000-essay dataset was used. Teachers were guided to use the features and settings discussed in this paper and encouraged to explore the feature and statistical options on their own to see if they could improve the model.

Workshop Observations and Focus Group

I paused at various stages during the focus group to allow participants to ask questions or make comments. I held an optional focus group immediately after the workshop for any participants who were willing to share their immediate perceptions of the software. Focus groups are often conducted in qualitative research and can lead to a more natural and relaxed environment for participants to reflect and share their thoughts freely. The workshop and focus group were held on Zoom, with audio transcripts collected through Zoom's auto transcribe tool. After the workshops, I carefully reviewed the transcripts for accuracy.

Individual Semi-Structured Interviews

The final phase for the participants involved semi-structure interviews intended to assess how the workshop changed participants' perceptions. A set of questions akin to those used in the initial survey was employed to evaluate shifts in teachers' confidence, attitudes, and views concerning the use of AES. Mirroring the workshop format, the interviews were conducted through Zoom, with audio transcripts being gathered. Participants who were unable to attend the interviews were given the opportunity to respond to the questions through email.

Data Analysis

Data analysis was conducted to address the research questions:

RQ1. Are teachers interested in using AES to enhance teaching and learning? Why or why not?

RQ2. How easily can teachers learn to use LightSide?

RQ3. What challenges did teachers face in learning to use the platform?

RQ4. How do trained teachers intend to use LightSide to enhance teaching and learning?

RQ5. Did the LightSide workshop influence teachers' perceptions about AES?

All data from the initial surveys, workshop and focus group transcriptions, and semi-structured interviews were coded using grounded theory to identify patterns in the data before applying Epistemic Network Analysis (ENA) to identify and visualize patterns in the data. The use of grounded theory and ENA together provides a comprehensive and nuanced analysis of the data.

Coding Procedures

During the transcription process, I removed all identifying information and assigned each participant a pseudonym. I segmented the data into lines of talk, with each line consisting of a natural sentence, pause, or turn of talk, and entered into a "text" column of a spreadsheet, as shown in Figure 3.17. ENA constructs networks based on units, conversations, stanzas, and codes. Units refer to people or groups, identified by "username" in this dataset. The code columns are the columns in vertical text in Figure 3.17. The codes make up the epistemic frame elements, or the nodes of the network model (Shaffer et al., 2016). Conversations, identified by "activity" here, consist of collections of text lines, usually with different time segments, activities, or steps of a process. Stanzas are lines that are related to one another. The ENA Webkit tool creates moving stanzas based on conversations within a group during the same activity on the same date. In Figure 3.17, Chris has three lines of talk:

- I've tinkered with ETS feedback through Turnitin.
- The disadvantage is that the feedback is often random and inaccurate.
- I want to learn more because I think the technology isn't ready yet, but maybe it's on the near horizon.

While this includes three separate lines of talk, the ideas are closely related. “The disadvantage” in the second line refers to the feedback tool Chris mentioned in the first line. In ENA analysis, the co-occurrence of elements in a stanza represents cognitive connections.

The relations among objects in the data is identified through the generation of adjacency matrices, with each matrix representing the co-occurrence of codes in stanzas. A binary summation of each code was used to indicate whether a code that appeared at least once in a stanza. Binary summation is appropriate for this dataset because there is no basis for assuming that a participant who mentions something twice as frequently necessarily comprehends or believes it to be twice as significant.

Figure 3.17

Excerpt of the Coded Log File

ACTIVITY	GROUP	DATE	USER NAME	TEXT	fear	doubt_uncertainty	difficulty	ease	AI_experience	AI_usefulness	AES_experience	AES_usefulness	curiosity	willingness	resistance	relevance
pre-survey	1	1/19/2023	Rae	I use learning analytics on Canvas to analyze student activity, see which resources are being used, identify students who are falling behind, etc.	0	0	0	0	1	1	0	0	0	0	0	0
pre-survey	1	1/19/2023	Rae	I used the built-in correction tools in turnitin.com	0	0	0	0	0	0	1	1	0	0	0	0
pre-survey	1	1/19/2023	Rae	I would like to find a scoring tool that gives students accurate results and provides teachers with a quick breakdown of students' problem	0	0	0	0	0	0	0	0	0	1	0	0
pre-survey	1	1/20/2023	Chris	I've tinkered with the ETS feedback available through turnitin.	0	0	0	0	0	0	1	0	1	0	0	0
pre-survey	1	1/20/2023	Chris	The disadvantage is that the feedback is often random and inaccurate.	0	1	0	0	0	0	0	0	0	0	0	0
pre-survey	1	1/20/2023	Chris	I want to learn more because I think the technology isn't ready yet, but maybe it's on the near horizon.	0	0	0	0	0	0	0	0	1	1	0	0
pre-survey	2	1/20/2023	Andy	One pro might be students getting faster responses on their essays.	0	0	0	0	0	0	0	1	0	0	0	0
pre-survey	2	1/20/2023	Andy	I worry teachers will rely too heavily on automated scores.	0	1	0	0	0	0	0	0	0	0	1	0

Coding Scheme

To address RQ1, which examines teachers’ interest in using AES to enhance teaching and learning, it was helpful to determine whether teachers have had prior experience with AES

and whether they consider it to be useful. Given that AES is an application of artificial intelligence and machine learning, this research question also aims to gauge the participants familiarity with AI and AES and their perceived usefulness in the context of education. The coding scheme for this research question included the following categories: *AI experience, AI usefulness, AES experience, AES usefulness, willingness* and *curiosity,* as shown in Table 3.2.

RQ2 and RQ3 are closely interrelated and aim to explore the usability of LightSide. RQ2 examines the ease of learning to use LightSide, with the aim of assessing the platform's accessibility for teachers who may not have a technical background or experience using similar software. RQ3 provides valuable insight into the areas where the platform may need improvement, such as areas in which the platform interface is not intuitive, and it captures some of the complexity of the platform's features. The coding category for these questions is *difficulty* (see Table 3.2).

R4Q seeks to provide insight into how LightSide can be effectively integrated into existing instructional practices and curricula. The coding category is *willingness*, which reflects the degree to which teachers are willing and ready to adopt LightSide and incorporate it into their teaching practice. The *AES_usefulness* code, used to measure RQ1, can also provide insight into the willingness of teachers to adopt LightSide, as it informs the extent to which teachers are likely to view LightSide as a valuable tool.

The final research question of this study aims to investigate the impact of the LightSide workshop on teachers' perceptions of AES. Coding for perceptions included *fear, uncertainty, curiosity, willingness, resistance,* and *relevance* (see Table 3.2). The *fear* category identified teachers' concerns about being replaced by technology, as well as other fears about the

negative effects of AES. *Uncertainty* measured teachers' doubts about any potential benefits of AES for learning. *Curiosity,* also used in RQ1, examined teachers' interests in learning more about the technology. *Willingness*, which was also used to address RQ4, looked at teachers' openness to learning more about AES, collaborating with other teachers, and using the technology in their own teaching practice. *Resistance* examined whether teachers felt opposed to adopting LightSide or other AES technologies. *Relevance* sought to determine the extent to which teachers felt compelled to stay up to date with technology to remain relevant in the teaching profession. Teachers were asked similar questions pertaining to the perception categories before and after the workshop. ENA was then used to identify changes in the connections between different coding categories, indicating any shifts in teacher's perceptions.

Table 3.2

Codes

Code	Definition	Example
AI Experience	Has experience using AI tools	I use learning analytics all the time to evaluate both individual students and whole classes
AI Usefulness	Considers AI to be useful for language teaching and learning	The data provided by Turnitin, Canvas, MyEnglishLab, Kahoot!, etc. informs my teaching practice
AES Experience	Has experience using AES tools	I've tried a multitude of auto-grading software applications.
AES Usefulness	Considers AES to be useful for language teaching and learning	I see it as a great learning tool for students.
Fear	Conveys fear that AI and AES will have negative effects on teaching and learning	Frankly, I'm worried about the future of my profession. I see automated essay scoring as a threat to teachers.
Uncertainty	Conveys uncertainty or doubts AI and AES will be beneficial for teaching and learning	I'm not convinced this is a good thing.

(table continues)

Code	Definition	Example
Difficulty	Experienced technical or other difficulties or is concerned about potential difficulties	I think more training will definitely need to be done before this could be used.
Resistance	Expresses unwillingness to adopt AI and AES technologies for own teaching practice	I don't have the stamina or ambition that I once had for our field and do not see automated systems as impacting my teaching whatsoever.
Curiosity	Expresses curiosity about AI and AES technologies	I'm curious about how it works.
Willingness	Shows willingness to learn more about the technologies, collaborate with other teachers, and use the technologies in own teaching practice	I expect AES will continue developing and improving and that it will become an integral part of what we do as writing instructors.
Relevance	Expresses desire to stay up to date in technology in order to remain relevant in the teaching profession	I do believe educators who do not keep up with new technologies will quickly become irrelevant.

Following the transcription of the collected data, it was segmented and organized into a spreadsheet for ease of analysis. Each segment was examined and assigned appropriate codes based on the categories mentioned earlier. This approach facilitated the identification of patterns, themes, and insights within the data. To ensure accuracy and reliability in the coding process, a qualitative researcher reviewed the transcriptions and codes, and a second rater independently coded the data to ensure reliability.

Processing the Data

ENA is a research method that uses network models to quantify relationships in coded qualitative data (Shaffer et al., 2016). The theory underlying ENA is based on the learning science theory of epistemic frames, which characterizes communities of practice in terms of the associations among knowledge, skills, perceptions, and other cognitive elements used to solve

complex problems (Shaffer, 2004). In this study, the community of practice is comprised of experience college ESL teachers who share the challenge of using AES in teaching and learning.

The coded data was uploaded to the ENA Webkit, a web-based application developed by Shaffer (2016). In epistemic networks, the connections among nodes represent the co-occurrence of codes. The network is weighted, with darker, thicker lines indicating stronger connections while lighter, thinner lines represent weaker connections. In this study, the lines represent connections among teachers' knowledge, experience, and perceptions of automated essay scoring. The connections are visualized by the first *(x)* and second *(y)* dimensions, in which *x* and *y* represent the greatest variation in data. Visualizations were created to measure the strength of association between codes at three different times: before, during and after the LightSide workshop. The visualizations enable more accessible exploration and insightful understanding of the data.

CHAPTER 4

FINDINGS

The purpose of this study was to examine ESL teachers' interest in and perceptions of AES, understand challenges in implementing AES, and explore potential uses for AES in ESL education. Eighteen teachers participated in the study by completing a pre-workshop survey, attending a two-hour workshop to learn how to build their own AES model using LightSide, and participating in a post-workshop interview. Female participants slightly outnumbered males. Nearly all participants were at least 40 years old. More than 60% were age 50 or older. Table 4.1 shows the gender and age distribution for the participants.

Table 4.1

Frequency Distribution of Gender and Age

Characteristic		Frequency	%
Gender	Female	10	55.6
	Male	8	44.4
Age	30-39	1	5.6
	40-49	6	33.3
	50-59	6	33.3
	>60	5	27.8

Most of the participants taught ESL in a community college or university, while one taught in an adult literacy center. All participants had taught ESL for at least six years, with 89% having more than a decade of teaching experience. The distribution of years of teaching experience is shown in Table 4.2.

When asked whether they agreed their technology expertise was high, ten participants agree, six neither agreed nor disagreed, and only one disagreed. When asked how they

integrate technology in writing classes, all participants indicated using a learning management system. Others had a wide range of experience implementing technology, as shown in Table 4.3.

Table 4.2

Frequency Distribution of Setting and Years of Teaching

Characteristic		Frequency	%
Teaching Location	Community college	14	77.8
	University	3	16.7
	Adult literacy center	1	5.6
Years of Teaching	6-10	2	11.1
	11 or more	16	88.9

Table 4.3

Technology Integration in the Classroom

Technology	Number of Participants
LMS	18
Active Board	5
Videoconferencing	16
Automated grading software	6
Plagiarism detection software	16
Autocorrection software	7
Combined media, such as digital storytelling	7
Collaboration via cloud platforms	11
Social media or blogs	6
VR or AR	3

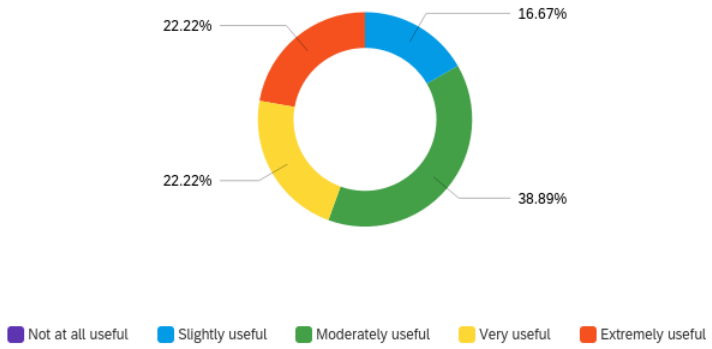
Teacher Interest in AES

The first research question sought to gauge teachers' interest in AES. Arguably, all

participants demonstrated a certain level of interest in using AES simply by voluntarily attending the LightSide workshop, which provided no compensation or professional development credit. Their participation in the workshop stemmed from genuine curiosity and interest in the topic. In the pre-survey, participants responded to Likert-type questions pertaining to the usefulness of ML, AI, and LA for teachers. Figure 4.1 reveals that nearly half of the participants considered these tools as either "very useful" or "extremely useful," while all participants rated them as at least "slightly useful." In describing prior experience with AI, ML, and LA, some participants mentioned using chatbots, plagiarism detection software, and LMS analytics.

Figure 4.1

AI, ML, and LA Usefulness Results

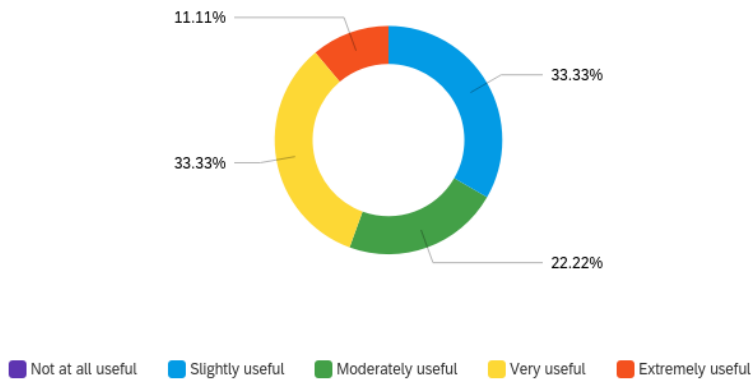


When inquired about the usefulness of AES specifically, approximately 45% of the participants evaluated it as "very useful" or "extremely useful," as shown in Figure 4.2. Consistently, every participant considered AES to be at least "slightly useful." Almost 50% of the participants indicated that they had prior experience implementing AES in their teaching practices. Among the AES systems mentioned were a diverse range of tools, including

SmarMarq, which is designed to assess and provide feedback on students' writing; Grammarly, a popular grammar-checking assistance tool; e-rater, a scoring engine developed by ETS; turnitin.com, a plagiarism detection software; and built-in tools integrated within learning management systems like Canvas.

Figure 4.2

AES Usefulness Results



Based on analysis of participants' surveys, focus group transcripts, and interview transcripts, interest in AES can be grouped into four main themes: (1) improving grading accuracy, (2) saving teachers' time, (3) delivering prompt results to students, and (4) facilitating self-directed learning. The following examples highlight these themes.

Table 4.4

Structure of Findings for Research Question 1

Themes	Subthemes
Theme 1: Improving grading accuracy	• Precision • Objectivity • Consistency • Calibration

(table continues)

Themes	Subthemes
Theme 2: Conserving time	• Reduced workload and stress • Increased focus on instruction • Opportunities for professional development
Theme 3: Delivering prompt results	• Relevance and retention • Facilitates revision • Increased motivation
Theme 4: Facilitating self-directed learning	• Self-assessment • Repeated practice • Personalized learning • Reduced anxiety

Theme 1: Grading Accuracy

Teachers expressed interest in having tools that yield precise results. AES systems can detect subtle patterns and linguistic features that may be difficult for human graders to notice, especially when teachers are required to grade a large number of essays in a short amount of time. Under such circumstances, AES systems may provide a more precise assessment of writing quality. The following comments from the pre-workshop survey illustrate this point.

> *Ben: I would like to find a scoring tool that gives students accurate results and provides teachers with a quick breakdown of students' problem areas.*

> *Andy: Robograding can detect patterns that work well in the aggregate over massive bodies of text.*

Other teachers mentioned that their desire for a tool that could confirm the consistency of their grades. Whereas human graders are subject to fatigue and personal biases, AES systems are objective, applying the same set of evaluation criteria to each essay and ensuring that every submission is assess consistently.

> *Henry: I use automated essay scoring to supplement my grading, as a way to sort of check myself.*

Rae: I'll probably use it as a second grader. I'm not really that confident that my own grades are consistent enough. I like the idea of a machine checking my scores.

Hannah: It will help me build confidence in my own grading. I always worry that I'm a bit too harsh. With essays, it's challenging to be consistent.

Similarly, teachers reported the need for better calibration of grading practices among multiple teachers or graders. By providing a benchmark against which human graders can compare their evaluations, AES systems can help identify inconsistencies and discrepancies within a program or department, leading to greater interrater reliability. Two teachers discussed this advantage during the focus group immediately after their workshop.

Rae: Just to standardize my own grades, I think this is very useful. It's probably even more use to standardize grades among all teachers, you know, get us all on the same page.

Cara: Right. You know how some teachers have a reputation as being a tough grader. We could really use something like this to prove it. Are they really too tough? This could give us answers.

Theme 2: Conserving Time

Several teachers mentioned their interest in utilizing AES to save time. Some mentioned that decreasing the time required for grading essays would allow teachers to concentrate more on pedagogy and instruction. In the pre-workshop survey, two teachers considered the decreased workload for teachers to be one of the biggest advantages of AES.

Tom: The obvious pro is that it makes teachers' lives easier.

Grading essays can be time-consuming and mentally exhausting, especially for teachers managing large classes. By automating the process, teachers can experience a reduced workload, leading to less stress and potential burnout. During the focus group, one teacher discussed the advantage of reducing teachers' workload, especially for placement tests.

Kate: We have thousands of essays to score for placement, and some of those students don't even end up studying in our program … we want to get it right because we don't want to place those students in the wrong level, but we also don't want to kill ourselves grading hundreds and hundreds of essays of students we'll never see.

Some participants noted that a reduced workload could allow them to increase their focus on instruction. Reducing the time spent on grading means teachers can allocate more time and energy to lesson planning, classroom instruction, and support for students, leading to improved learning outcomes and a better learning experience for students. More time would also allow more opportunities for teachers to engage in professional development activities and collaboration with colleagues.

Theme 3: Prompt Results

Teachers recognize the benefits of prompt feedback for students and are consequently interested in employing AES to generate more rapid results for their learners.

Andy: One pro might be students getting faster responses on their essays.

As some participants observed, providing feedback while the material is still fresh in their students' minds helps students understand the connection between their work and the feedback. The feedback is still relevant and can thus enhance retention. As one teacher noted, prompt feedback also facilitates the essay revision process because it allows students to apply the suggested improvements and revise their work more efficiently.

Emily: Students who receive quick feedback are more likely to revise their essays carefully. If we wait too long, students will have forgotten about their topic or moved on to other assignments.

Another teacher pointed out that prompt feedback can boost students' motivation, as they can see immediate results of their efforts.

Sam: I think it motivates them to see their progress.

Overall, prompt feedback is a crucial aspect of the learning process, as it helps students understand their performance, stay engaged, and develop better writing skills.

Theme 4: Self-Directed Learning

Teachers showed great interest in tools that facilitate self-study for their students. With this in mind, teachers seek tools that students can use independently to track their progress, improve their essays, and develop into more independent, self-reliant writers.

> *Hannah: I am currently exploring how to use chatbots to engage our students in self-study activities and possibly assist teachers in their grading.*

> *Ben: I would like students to use an automated system as a type of formative assessment to guide them on their journeys as writers.*

Self-assessment also allows for repeated practice. As mentioned earlier, it facilitates the revision process, which would allow students to submit multiple drafts of their essays and receive feedback on each attempt. This iterative process helps students learn from their mistakes and refine their writing skills.

> *Kim: I see an opportunity to change the way writing is taught, using automated assessment as the backbone of the revision, editing, and feedback cycle.*

Other instructions spoke to the value of personalized learning, which empowers students to take charge of their learning journey and work on specific problem areas.

> *Tom: I think if we could create a nice, neat list of each student's strengths and weaknesses, in terms of features, I think that could be very good for personalizing instruction. I can hone in on those problem areas.*

> *Sam: Automated essay scoring will provide insight to what's going on in our student's writing. I can show it to the students so they can track their progress.*

Lily, a teacher who is not a native English speaker, discussed how AES can reduce

student anxiety associated with human grading. The reduction in anxiety can create a more conducive learning environment and empower students to take charge of their own learning.

> *Lily: Some students wanted to practice their writing skills, but they were a little uncomfortable showing me their essays because they knew that they were not doing a great job. So if I tell them this is not going to be a worry or a concern that they should have because their essay is going to be graded by a machine or LightSide software or something like that, I think more students will feel comfortable.*

Overall, teachers report that AES supports self-directed learning by providing immediate feedback that can be used for revising essays, helping teachers better understand their students' writing in order to personalize instruction, encouraging self-assessment and reflection, and fostering a more comfortable and efficient learning experience.

Reluctance

It is important to acknowledge that although all participants exhibited some level of interest in AES, not everyone was entirely enthusiastic about adopting this technology. Some conveyed reservations about implementing AES, as demonstrated by the following examples from the pre-survey.

> *Tom: I have zero expectations for this technology and hope that it is only a passing fad.*

> *Kate: At this stage in my career, I do not expect to transition to automated scoring. I will continue using my own tried and tested scoring system of reading and evaluating my students' essays myself.*

Experience and Interest Connection

ENA results from the pre-workshop survey show a strong connection between prior experience with AES programs and perceptions of usefulness, which is connected to a willingness to learn more, versus resistance, illustrated in Figure 4.3. This ENA model included

the following codes: curiosity, willingness, resistance, AES_usefulness, AES_experience, AI_usefulness and AI_experience. Conversations were defined as all lines of data associated with a single value of "activity." In Figure 4.3, the conversation consisted of all the lines associated with "activity" and pre-survey. The ENA model normalized the networks for all units of analysis before they were subjected to a dimensional reduction, which accounts for the fact that different units of analysis may have different amounts of coded lines in the data. For the dimensional reduction, a singular value decomposition was used, which producing dimensions that maximize the variance explained by each dimension. The nodes correspond to the codes, and edges reflect the relative frequency of co-occurrence, or connection, between two codes. The result is two coordinated representations for each unit of analysis: (1) a plotted point, which represents the location of that unit's network in the low-dimensional projected space, and (2) a weighted network graph. The positions of the network graph nodes are fixed, and those positions are determined by an optimization routine that minimizes the difference between the plotted points and their corresponding network centroids. Because of this co-registration of network graphs and projected space, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. This model had co-registration correlations of 0.94 (Pearson) and 0.94 (Spearman) for the first dimension and co-registration correlations of 0.91 (Pearson) and 0.92 (Spearman) for the second. These measures indicate that there is a strong goodness of fit between the visualization and the original model.

When asked about their initial perceptions during the workshop, participants expressed interest in using LightSide, with comments such as "It looks like a cool program," and "I'm

intrigued." Figure 4.4 shows participants made stronger connections between curiosity, willingness, and AES usefulness during the workshop.

Figure 4.3

Epistemic Frame for Interest in AES before the Workshop

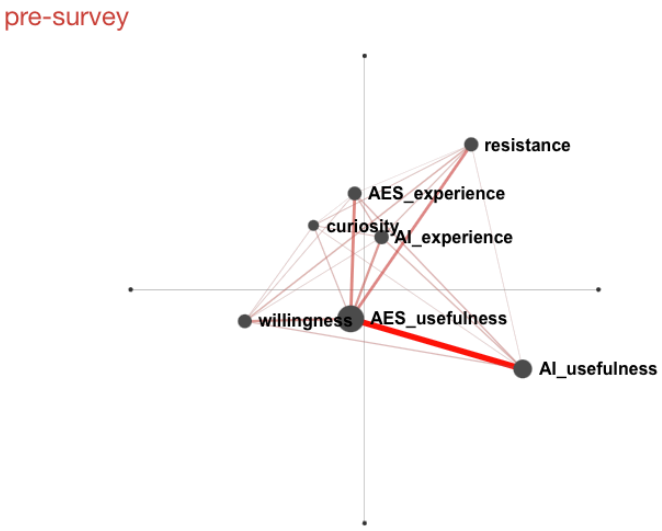
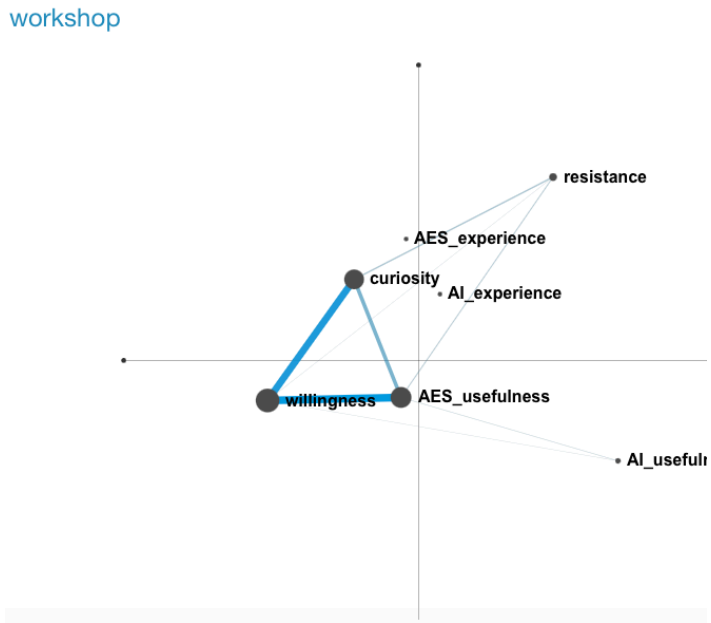


Figure 4.4

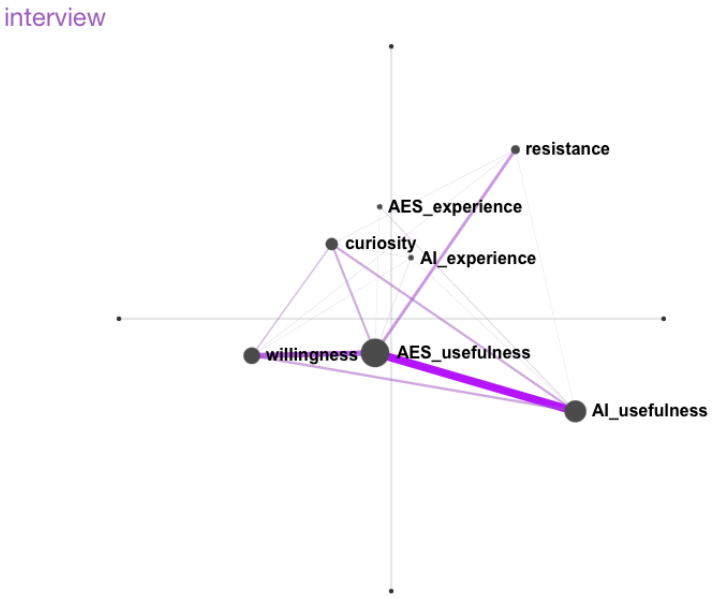
Epistemic Frame for Interest during Workshop



After the workshop, participants continued to make connections between usefulness, curiosity, and willingness, as illustrated in Figure 4.5.

Figure 4.5

ENA Analysis of Interest in After Workshop



ENA results revealed significant differences between teachers' interest in AES before, during, and after participating in the LightSide workshop. Figure 4.6 displays these differences in the form of an ENA means plot. The red dots represent teachers' interest in AES before the workshop, based on data from the pre-workshop survey. The blue dots represent data collected during the workshop focus group. The purple dots indicate participants' interest after the workshop, derived from data collected during the interviews. ENA allows for the comparison of units of analysis in several ways, such as plotted point positions, individual networks, mean plotted point positions, and mean networks, which average the connection weights across individual networks. Network difference graphs can also be employed to compare networks by calculating the difference in connection weights between two networks.

Figure 4.6

ENA Comparison of Interest

A two-sample t-test assuming unequal variance was conducted to compare pre-survey and interview data along both the X and Y axes. Along the X-axis, the pre-survey and interview groups showed no statistically significant difference at an alpha=0.05 level. However, along the Y-axis, there was a statistically significant difference between the pre-survey and interview groups at the same alpha level. ENA results demonstrated that the LightSide workshop significantly influenced the teachers' perspectives on AES. There were noticeable differences in their views before and after participating in the workshop, as shown by the significant difference along the Y-axis in the ENA plot.

Table 4.5

Comparison of Pre-Workshop Survey and Post-Workshop Interview Variance

	X Axis SVD1		Y Axis SVD2	
Mean	0.44	0.27	0.15	-.30
Standard Deviation	0.50	0.65	0.62	0.66
N	18	18	18	18

Note: SVD1: t(32.02) = -0.88, p = 0.39, Cohen's d = 0.29. SVD2: t(33.84) = 2.10, p = 0.04, Cohen's d = 0.70

User-Friendliness

The second research question examined how easily teachers could learn to use LightSide. Interestingly, this concern was mentioned only once in the pre-surveys, suggesting the participants likely anticipated the software to be intuitive and user-friendly. One participant noted in his pre-survey that he would probably struggle with learning to use the software. Other participants did not express concerns about a learning curve. However, during the workshop, only a few participants found the software to be easily navigable. By the end of the training, half of the participants reported that they felt they would require additional training beyond the two-hour workshop. This contradicts the claims made by LightSide's creators: "We've built a tool that lets you hit the ground running with your data, putting as much of the research workflow as possible in an easy, point-and-click interface" (Mayfield et al., 2014, p. 1). Not only were the teachers *not* able to hit the ground running, they also remained unconfident in using the software even after a two-hour workshop that guided them step-by-step through the process of understanding the interface, uploading data, training and building an AES model, and interpreting the results.

In the interviews conducted one to two weeks after the workshop, ten out of the

eighteen participants managed to use the software, although the majority encountered some technical issues. Eight participants did not even attempt to use the software, as they deemed it too challenging and intimidating. In other words, ease of use presented a problem both during the workshop and as participants attempted to practice using the software on their own. Specific issues are discussed in the next section.

Challenges

While the LightSide developers claimed to have built an intuitive, user-friendly platform, it is still a text-mining platform that involves technical complexity. Participants experienced several challenges before, during, and after the workshop. The challenges can be grouped into three themes: technical issues, data preparation, and interface design.

Technical Issues

Simply setting up LightSide and preparing for its use involves opening a terminal window and changing the memory settings on the user's computer and installing a supported version of Java. These initial steps alone frustrated several participants. Before the workshop took place, I reached out to each participant to help set up LightSide on the participant's device. In some cases, this step took over an hour. LightSide does not perform consistently on different computers. Some participants complained that LightSide's documentation was not clear or comprehensive enough. These frustrations persisted during and after the training. One participant experienced a software crash during the workshop.

> *Cara: It was okay, but then it just sort of crashed, the software. It just stopped after I loaded the dataset.*

Another participant reported having to reinstall Java on his system again before he was able to use LightSide after the workshop. Another complained that he could not open or use LightSide at all after the workshop. Nearly all participants reported the need for more training.

Data Collection and Preparation

Beyond technical problems, participants seemed overwhelmed by LightSide's customization requirements. While customization is one of the platform's key advantages, participants expressed uncertainty about how to collect enough of their own data in order to build and train an effective model.

> *Alice: For it to be useful, I think, we'd probably need a couple hundred essays for each score, and I just don't know how realistic that is.*

Other participants felt overwhelmed with the task of formatting the data in a spreadsheet with data columns.

> *Scott: It would be great if we could just dump all the Word docs in a folder and then upload that into LightSide somehow. Putting it all in a spreadsheet just feels like a big waste of time. Makes me wonder if we're really saving time at all.*

> *Kate: How do we create a base, a model you call it? How do we build that to score our own essays? who's going to do all that work? It seems incredibly labor intensive, creating this kind of spreadsheet.*

In addition to collecting and formatting data, users need to train their own models, which can also be time-consuming and involves a lot of trial and error.

Interface Design

While LightSide aims to provide an easy, point-and-click interface, participants still found the interface to be unintuitive and confusing. With multiple tabs and options, some participants found it difficult to navigate and understand the workflow. For some, the

terminology was difficult.

> *Rae: I would like something that can show us results in terms of grammar terms instead of features. I know LightSide provides that information, but it uses the language of computational linguists instead of applied linguists. I wish it had more teacher friendly language.*
>
> *Cara: The terms are hard to understand. So features are parts of speech? That's what I can't quite wrap my brain around. I guess I just don't get it.*

Others were confused by the algorithms involved in training a model.

> *Andy: I've actually studied statistics, but, I mean, I don't know that any ESL teacher can really use this.*

Some felt intimidated by the entire platform. One participant said that he was able to open the platform and follow along during the workshop, but he could not understand what he was viewing. Table 4.6 summarizes participants' challenges with the platform.

Table 4.6

Structure of Findings for Research Questions 2 and 3

Themes	Subthemes
Theme 1: Technical issues	• Changing memory settings • Installing Java • Running software • Program crashes • Lack of technical support
Theme 2: Data collection and preparation	• Need for large corpus • Formatting data
Theme 3: Interface design	• Understanding statistics • Confusing NLP terminology • Making sense of results

Potential Use

The fourth research question sought to understand how teachers intended to use LightSide to enhance teaching and learning. During the post-workshop interviews, teachers

were asked how they could use LightSide in their instruction. Two teachers said they had no plans for using LightSide again because they did not find it helpful. Two other teachers expressed some doubt about using LightSide because of its ability to produce a single score rather than detailed feedback.

> *Andy: The crucial problem as I see it is the inability of the system to provide accurate and useful feedback on the level of a single piece of writing, non-statistical judgment as it were. Without accurate feedback, what good is this system really providing for the students who are doing the writing as part of a learning process? My goal is to help students with the content of their writing, the ideas, and with the writing process as a whole. Statistics do not help my students.*

Joel, who is also the testing coordinator for his program, said he would consider using LightSide for testing purposes but not for classroom instruction:

> *I think it will be helpful to me as the writing assessment coordinator but probably not as an instructor. I don't see how I could use it in my day-to-day classes.*

A sixth teacher, Scott, expressed doubt about using LightSide for instructional purposes:

> *Automated scoring is going to be very useful, but I'm still not sure about LightSide. I mean, it's just not easy to use. I think maybe if you have a strong background in statistics, but for me, it's not very easy. I want to try it, and I really want to figure out something we can do with it, but I don't know. I think it's going to be good for standardizing our grading, like we all keep talking about, but two big problems. It doesn't give us a clear breakdown of the score, like a rubric I mean. I know we can see the features or whatever, but most of us don't even know what we're looking at with that. Features are just numbers. More statistics. And two, we have to reformat all our essays. It would be great if we could just dump all the Word docs in a folder and then upload that into LightSide somehow. Putting it all in a spreadsheet just feels like a big waste of time. Makes me wonder if we're really saving time at all. So I think we need something a lot more automated than LightSide.*

The remaining 12 participants expressed their intention to incorporate LightSide into their instruction. These participants identified various ways in which they would use the tool, highlighting its potential benefits and contributions to their teaching practice. Usage themes

included grading assistance, grade calibration, keeping up with technology, gaining insight into students' writing, and AI detection.

Grading Assistance

Teachers acknowledged the potential of LightSide to streamline the grading process by providing accurate scores for student essays:

> *Beth: I was impressed with how efficiently Lightside can provide the information we need. Clearly it can reduce the time and effort of the teacher and provide an instant feedback for the teacher and student. I'll use it as an assistant in grading students essays in the writing class, and preparation of the English writing test in TOEFL and IELTS. I just wish it could give more detailed feedback of essays such as actual grammar mistakes, topic relevance, unity and coherence.*

> *Kim: Students need to write every single day, but teachers don't have the time or energy to read and grade papers every single day. Automated scoring can give us a break, while ensuring students continue getting the practice they need.*

By using LightSide as a support tool, teachers can reduce the time spent on manual grading, allowing them to focus instead on other aspects of teaching and student support.

Calibration

Some teachers intend to use LightSide to calibrate their grading practices, ensuring consistency and fairness in evaluating students' work:

> *Lily: It is helpful because I will gain confidence in my grading. Sometimes I don't know how to justify my marks, A, B, C, or D. The difference is not always crystal clear. I want to have this tool supervise my marks.*

> *Kate: It will be enormously helpful and will take pressure off teachers who are responsible for grading not just their own students' work, but also placement and exit tests, as well as student portfolios. I like the idea of an objective grader. What is more objective than a machine?*

By comparing their own grading with LightSide's automated scores, teachers can

identify potential biases or discrepancies and can adjust their assessment practices accordingly.

Relevance

Some participants expressed a desire to stay current with technological advancements in education. At the time of this study, AI is a very hot topic in education. By incorporating LightSide into their teaching, teachers can explore the potential of AI-driven assessment tools and enhance their overall teaching practice through the integration of new technologies.

> *Cara: It is helping me stay current and gain a deeper sense of the technology. With the Chat GPT craze, AI is undoubtedly going to be yet another thing we have to add to our already overflowing toolbox. I honestly don't know how we can possibly keep up.*

Insight

Some participants noted that LightSide can help them analyze patterns and trends in their students' writing, offering valuable insights into students' strengths and weaknesses. By examining the extracted features and model outcomes, educators can gain a deeper understanding of their students' writing skills and tailor instruction to address specific areas for improvement.

> *Sam: Automated essay scoring will provide insight to what's going on in our students' writing. For many years, I have made it a practice to keep a spreadsheet, a log if you will, to track my students' grammar mistakes. So for example, I put the student's name in one column, and then I have columns for different types of mistakes. Subject-verb agreement, verb tense, run-on sentence, pronoun reference, and so on. Every time I grade a set of essays, I write the number of each mistake for each student.*

AI Detection

Study participants also recognized the potential of LightSide in detecting instances of AI-generated text in student essays. This, too, is a very hot topic in education at the time of this

study. As AI tools become more and more prevalent, it becomes increasingly important for educators to ensure the authenticity of their students' work. LightSide could serve as a valuable resource for identifying AI-generated content and promoting academic integrity.

> *Maria: It will most certainly be useful to instructors. I hope we can develop it as an AI detection tool. We know our students are already using AI to write essays. We need a solid system to detect whether an essay was written by a bot or a human. I believe LightSide has the capability, as far as I could tell from the workshop.*

Teacher Perceptions

The final research question in this study explores whether the LightSide workshop altered teachers' perceptions of AES. To address this question, a well-structured coding framework was established, encompassing categories, themes, and subthemes. Data collected from pre-workshop surveys, focus groups, and post-workshop interviews unveiled several key themes: fear, uncertainty, AI and AES usefulness, curiosity, willingness, resistance, and relevance.

Negative perceptions included fear, uncertainty, and resistance. Participants expressing fear voiced apprehensions about the potential for AES to replace their roles as educators or negatively impact the teaching and learning processes. These concerns often revolved around the loss of personal interaction between teachers and students and the possibility of overlooking unique student needs due to overreliance on technology, as can demonstrated by the following comments.

> *Tom: The con is that we are dehumanizing the teaching and learning process.*
>
> *Kim: I'm deeply concerned about the pitfalls, that teachers will rely too heavily on automated tools.*

Other teachers, concerned being replaced by machines, expressed fear of job loss.

Ben: The idea of robots replacing teachers in the classroom seems a dystopian outlook.

Maria: Frankly, I'm worried about the future of my profession. I see automated scoring as a threat to teachers.

Uncertainty and doubt emerged from concerns regarding the possible ineffectiveness or lack of benefits of AES in language education. Some participants questioned whether AES systems could accurately assess the complexities of language and the intricacies of student writing, as well as provide meaningful feedback for student growth.

Chris: The feedback is often random and inaccurate.

Lily: It's probably not as good as human raters in judging whether there is logical thinking and effective communication of ideas.

Beth: It focuses on too many errors, so the student may be overwhelmed.

Cara: I felt doubts about the ability of machine to properly and fairly assess my students.

In a similar vein, resistance manifested in participants who indicated they had no interest or desire to incorporate AES into their instruction. This reluctance often stemmed from a deep-rooted belief in the importance of human judgment in evaluating student work and the fear that technology could not adequately replace the intuition and expertise of an experienced teacher.

Tom: I have been teaching for over 30 years and have no plans to change my method of teaching and assessing students.

Kate: I will continue my own tried and tested scoring system of reading and evaluating my students' essays myself.

Some teachers exhibit resistance towards learning new tools, particularly if they are nearing retirement, as they may perceive the time and effort required to master these technologies as an unwarranted investment at this stage in their career.

Ben: I don't have the stamina or ambition that I once had for our field.

Henry: I'm on the road to retirement. [These tools] will not help me personally.

Scott, a relatively younger teacher, articulated his resistance to adopting LightSide, citing his belief that the rapidly evolving landscape of AES will likely yield newer and more advanced solutions in the near future. As a result, he feels hesitant to invest time and energy in learning to use a tool that may become outdated or obsolete soon, preferring instead to wait options that are more advanced and intuitive to emerge:

> *The dilemma is when something is getting better really fast, people don't want to adopt tech because it's getting better so fast that there's a sense a better one is coming, so why learn this one right now?*

On the other hand, curiosity-driven participants were keen to explore and learn more about AES. These participants viewed AES as an opportunity to expand their pedagogical toolkit and enhance their students' learning experiences.

> *Andy: I think it's useful to understand how things work because otherwise it's just kind of magic… I'm at a point in my life where I've got some curiosity, and I just want to see how it all works.*

Similarly, participants expressing willingness demonstrated enthusiasm to further their knowledge, collaborate with fellow educators, and integrate the technology into their teaching practices.

> *Kate: Recent articles on AI have piqued my interest in how these technologies can be applied in an effective and ethical manner.*

> *Hannah: I have used [Lightside] with my colleagues. We are creating our own set of data to train a new grading model.*

The theme of relevance emerged from participants' desires to stay current with technological advancements and maintain their professional relevance in the ever-evolving field

of education.

> *Beth: Our teaching tools are going to be advancing, just as the students' tools are, so it's going to be this race to be one step ahead.*
>
> *Hannah: If we hope to stay relevant in this field or any other, we must advance along with the technology.*

By examining these themes and their interplay, the study sheds light on the extent to which the LightSide workshop influenced teachers' perceptions of AES and its potential implications for language education. To compare teachers' perceptions, similar questions were asked in the pre-workshop survey and the post-workshop interview. An ENA analysis reveals the workshop seemed to have influenced teachers' perceptions. Figure 4.7 shows participants made more connections between fear and uncertainty before the workshop, whereas Figure 4.8 shows stronger connections between usefulness, curiosity, and willingness after the workshop.

Figure 4.7

Epistemic Frame of Perceptions Before Workshop

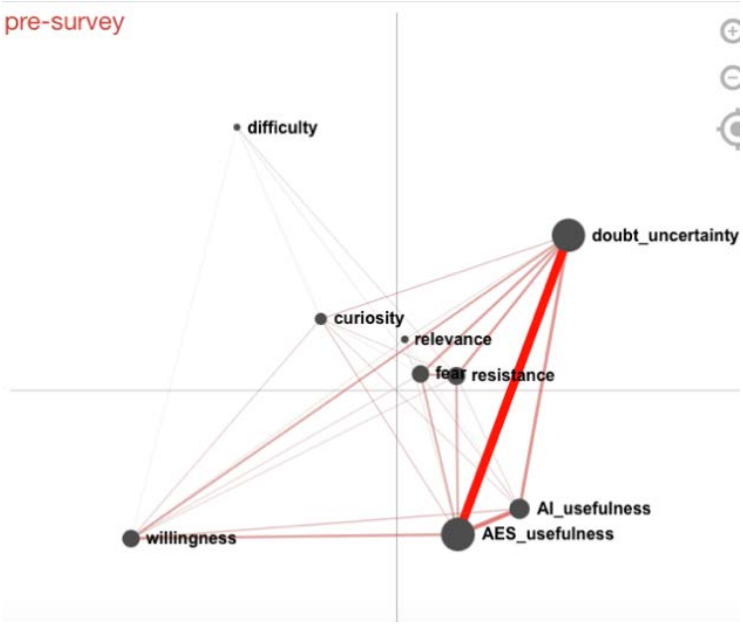
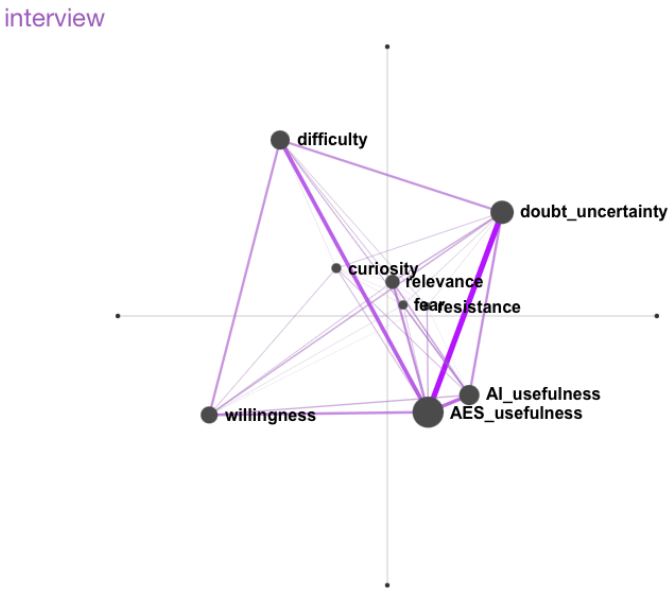


Figure 4.8

Epistemic Frame of Perceptions after Workshop



A comparative analysis is shown in Figure 4.8. Here, the ENA model included the following codes: curiosity, willingness, resistance, AES_usefulness, AI_usefulness, doubt_uncertainty, relevance, fear, and difficulty. Conversations were defined as all lines of data associated with a single value of "activity" (pre-workshop survey and post-workshop interview). This model had co-registration correlations of 0.94 (Pearson) and 0.96 (Spearman) for the first dimension and co-registration correlations of 0.94 (Pearson) and 0.94 (Spearman) for the second. These measures indicate that there is a strong goodness of fit between the visualization and the original model.

To test for differences, a two-sample t-test was applied, assuming unequal variance to the location of points in the projected ENA space for units in pre-survey and interview. Along the X axis, a two sample t test assuming unequal variance showed the pre-workshop survey (mean=-0.21, SD=0.67, N=18) was not statistically significantly different at the

alpha=0.05 level from the interview. Along the Y axis, a two sample t test assuming unequal variance showed the pre-survey (mean=0.33, SD=0.45, N=18) was statistically significantly different at the alpha=0.05 level from the interview (mean=-0.16, SD=0.44, N=18; t(34.00)= 3.32, p=0.00, Cohen's d=1.11). The differences in these relationships before and after the teachers participated in a workshop on AES, suggesting that the workshop might have influenced their views.

Figure 4.9

ENA Comparison of Perceptions Before and After Workshop

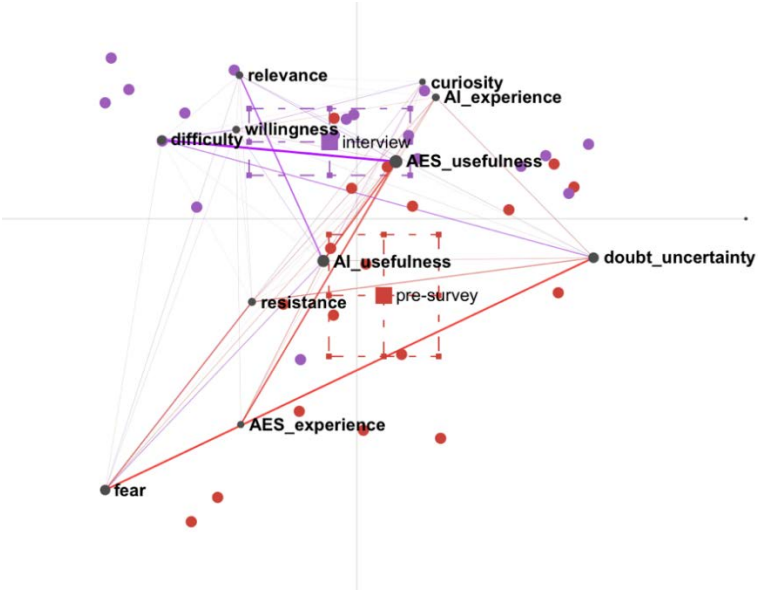


Table 4.7

Comparison of Variance in Perceptions

	X Axis SVD1		Y Axis SVD2	
Mean	0.53	0.10	-0.16	0.05
Standard Deviation	0.55	0.50	0.45	0.48
N	18	18	18	18

Note: SVD1: $t(33.76) = 2.46$, p = 0.02, Cohen's $d = 0.82$. SVD2: $t(33.84) = -1.29$, p = 0.21, Cohen's $d = 0.43$

Summary of Findings

This study aimed to answer five questions regarding teachers' interest in using AES and their experiences with the LightSide platform.

RQ1. Regarding teacher interest, the majority of teachers expressed interest in using AES to enhance teaching and learning. Their reasons included faster feedback, facilitating self-directed learning for students, and improving grading accuracy.

RQ2. Learning to use LightSide proved challenging for many teachers. While some participants initially expected the software to be intuitive and user-friendly, by the end of the two-hour workshop, half of them felt they would require additional training to confidently use the platform.

RQ3. Challenges teachers faced in learning to use LightSide included difficulty in navigating the software, the complexity of training and interpreting an AES model, and a lack of targeted, formative feedback alongside the generated scores.

RQ4. Trained teachers who expressed interest in using LightSide intended to utilize the platform for various purposes such as grading assistance, grade calibration, keeping up with technology, gaining insight into students' writing, and AI detection.

RQ5. The workshop significantly influenced teachers' perceptions about AES. The Epistemic Network Analysis (ENA) results showed noticeable differences in teachers' views before and after participating in the workshop, demonstrating the impact of the training on their perspectives.

CHAPTER 5

DISCUSSION

Although AES systems offer numerous advantages for language teaching and learning, many ESL teachers remain hesitant to implement this technology. Teachers have compelling reasons for this resistance. Some teachers view AES as a potential threat to the teacher-student relationship or believe that it undermines the importance of human judgment in evaluating written work (Shermis, 2014). Others argue that AES might not provide reliable and valid scores, as it could fail to capture the nuances and complexities of their students' language (Bennett & Zhang, 2016). Additionally, there are concerns that AES systems rely primarily on surface features, such as grammar and vocabulary, without adequately capturing the depth and quality of the content, rhetoric, critical thinking, and logical organization in an essay (Perelman, 2014).

While these are legitimate concerns, ESL teachers and students have much to gain from the implementation of AES, as discussed in this dissertation. Weigel (2014) suggested an effective way to overcome teacher resistance to technology is to expose teachers to tools. Though this study was limited in scope, the findings supported Weigel's claim. Teacher participants made stronger connections between AES and fear, doubt, uncertainty, and resistance before they attended a two-hour workshop to learn how to use LightSide. After the workshop, teachers expressed greater degrees of willingness to adopt AES and a desire to learn how to use AI and AES in order to remain relevant in their professions.

Introducing ESL teachers to LightSide helped alleviate fears, uncertainty, and resistance to AES, while engaging them in the ongoing discussion about AI and AES educational

applications. If ESL teachers do not attempt to understand this technology and engage in its usage, they risk being left out as developers continue to create better AES tools. LightSide serves as a valuable starting point, but it does not present a ready-made solution on its own. The following discussion explores the shortcomings of LightSide and proposes potential solutions.

Risk of a Weak Model

While the ability to customize models is one of LightSide's greatest strengths, it can also be a limitation for users who lack the necessary expertise to properly train and adjust models for their specific context. Building a well-performing model requires trial and error, which can be frustrating and time-consuming. Users who have no experience in building models might not know where to begin. Unlike commercial systems, LightSide does not come with pre-built models for assessing essays. Users need to train their own models using their own students' essays.

Collecting enough essays for an effective model might be challenging. Moreover, teachers need to be very careful to include essays that represent a range of language backgrounds and proficiency levels. Like any AES system, Lightside's models can be subject to biases inherent in the training data. If the training data does not represent the target student population very well, the model's performance could lead to inaccurate or unfair scoring.

Need for Training and Support

As previously noted, learning to use LightSide effectively requires time and effort. This was a consistent theme throughout this study. Participants consistently reported the need for

additional training and support to understand the software's capabilities and how to apply them to their specific use case. The learning curve proved to be especially steep for those with limited experience in computational linguistics or statistics. The technical knowledge required to use LightSide effectively make it challenging for novice users. Furthermore, as an open-source project, LightSide does not have the same level of support, updates, and maintenance as commercial AES systems, which leads to potential issues with compatibility, bug fixes, or new feature development.

Potential Solution: Communities of Practice and Special Interest Groups

A potential solution for the lack of support could be the forming of communities of practice (CoPs) and special interest groups (SIGs). Such groups would allow teachers to exchange knowledge and experiences related to the use of LightSide or other machine learning platforms, and discuss best practices and teaching techniques. This exchange of knowledge can help teachers overcome the learning curve associated with the software. A CoP or SIG could also facilitate collaboration among teachers, giving them opportunities to share resources, such as training datasets, model configurations, or feedback strategies, leading to more effective and efficient use of LightSide. Working together, teachers can expand their expertise in areas such as machine learning and text analysis. Armed with this information, ESL teachers can advocate for the use of AES tools in their institutions and engage in dialogue with developers to make improvements to existing tools and create new tools specifically designed to assess the writing of non-native English speakers.

Need for Feedback Supplement

LightSide uses interpretable model features that are based on psychometrics, which typically characterize writing elements such as lexical sophistication and coherence, with a focus on ensuring the defensibility of the model, or construct validity, which is accomplished by carefully selecting features (Mayfield, 2020). This approach is useful for training a scoring system that mimics a teacher's scores, but the emphasis is placed on construct validity rather than the ability of the model to offer practical writing suggestions based on the score. In addition to a score, students need targeted formative feedback to develop their writing skills. Automated scoring contributes value to the classroom; however, targeted formative feedback accompanying those scores is essential for developing writing proficiency.

On its own, LightSide fails to deliver meaningful feedback on student writing. It is a valuable tool, but it is not enough. Turnitin has adopted LightSide technology to create Turnitin Revision Assistant, joining similar programs such as TenMarks Writing, Grammarly, ETS Criterion, Pearson WriteToLearn, and Vantage MyAccess, to prioritize feedback in AES systems (Mayfield, 2020). Though extremely useful, such proprietary technology is often expensive and the exact approach and scoring algorithms vary greatly between systems. ChatGPT , a generative AI system, may offer a solution for providing feedback as a supplement to LightSide's score, though such systems come with their own risks and limitations.

Potential Solution: Chatbots

Chatbots, such as Chat GPT, could be a valuable supplement to LightSide in evaluating students' essays by addressing some of the limitations associated with AES systems. Chat GPT is a model developed by OpenAI, a research organization whose mission is "to ensure that

artificial general intelligence benefits all of humanity" (OpenAI, 2022). Based on patterns it learned from massive amounts of text data, Chat GPT uses deep learning techniques to produce human-like responses to a given question or prompt. The company website announced the release of Chat GPT at the end of the year in 2022 and invited users to try the research preview for free. As of the writing of this dissertation, Open AI has not publicly indicated how long the trial period will last, or what the cost of the program will be after the free trial period.

A potential application of Chat GPT is to provide it with a scoring rubric consisting of clear grading criteria and then ask it to evaluate and provide detailed feedback for a set of essays based on those criteria. Teachers can customize the feedback by supplying the scoring rubric and for requesting additional feedback for specific problems, such as grammar or vocabulary. While LightSide can generate a single score for an essay, Chat GPT can offer specific formative feedback, which can help students identify areas that need improvement and guide them in revising their work. Powered by AI, Chat GPT can potentially recognize and evaluate higher-order skills such as critical thinking, logical organization, and rhetorical strategies that might be challenging for LightSide to evaluate. Chat GPT can also engage students in a dialogue about their writing, asking questions and providing suggestions. This interactive approach can help students better understand the feedback and gain a deeper insight into their writing strengths and weaknesses.

The disadvantage to using Chat GPT is that teachers cannot see or adjust the algorithm and features Chat GPT uses for its scoring model. It might over or underemphasize certain features. Another potential risk of using Chat GPT is bias. Its training data is vast and may contain biases; as a result, it may unintentionally perpetuate those biases when assessing

essays, leading to unfair grading. As a generative system, Chat GPT is always evolving. It can produce non-facts and misinformation. Its performance depends on the size and diversity of its training data and the quality of its algorithms. As a result, Chat GPT may not always provide accurate or consistent grading. Another risk of using Chat GPT involves data and privacy concerns. Submitting students papers to an online AI system may raise data privacy issues, as sensitive student information could potentially be processed and stored by the system.

By supplementing LightSide with Chat GPT, teachers can provide a more comprehensive and interactive evaluation of students' essays. Table 5.1 summarizes what each tool can do, illustrating how the two programs may work well together.

Table 5.1

Comparison of LightSide and ChatGPT as Automated Scoring Systems

LightSide	ChatGPT
• Trained on user-provided data • Allows users to choose specific features and try different algorithms • Provides built-in features to test validity and reliability of the model	• Trained on big data • Evaluates essays according to a given scoring rubric • Provides detailed feedback on specific aspects of the essay, such as grammar, vocabulary, cohesion, etc.

Study Limitations

One notable limitation of this study is the relatively small sample size consisting of only 18 teacher participants. These participants were drawn from a rather homogenous demographic, with the majority being over the age of 40 and over half aged 50 or older. All of the teachers involved were highly experienced, boasting at least a decade of ESL teaching experience. All but one participant had a background in higher education settings, specifically

community colleges or universities. A more comprehensive study exploring the perceptions and usage of AES in ESL instruction would be useful, one that incorporates a diverse range of teachers of varying ages, educational settings, and levels of experience. By broadening the scope of the study, it would be possible to gain a more accurate and nuanced understanding of how different ESL teachers perceive and utilize AES.

Moreover, this study focused exclusively on ESL, limiting its applicability to the broader field of language education. Conducting comparative studies that explore the use of AES in other language education contexts, i.e., teaching languages other than English, could provide valuable insights into the efficacy and relevance of AES across various language teaching scenarios. Such research would contribute to a more comprehensive understanding of the role and potential of AES in language education as a whole.

Conclusions

By exploring the usability of LightSide in this study, I aimed to gain a better understanding of the factors that influence ESL teachers' ability to use text analysis tools and to identify areas where improvements can be made to enhance the platform's usability. I discovered that teachers were highly interested in using LightSide both as a way to enhance teaching and learning and to stay abreast of the rapidly advancing AI technology in order to maintain relevance in their profession. However, teachers reported the learning curve associated with using LightSide was rather steep. I hope this research will contribute to the development of more accessible text analysis tools that can support teaching and learning in second language classrooms.

APPENDIX A

PRE-WORKSHOP SURVEY

Exploring Uses of Automated Essay Scoring for Learner English

Consent TITLE OF RESEARCH STUDY: Exploring Uses of Automated Essay Scoring for ESL Teachers: Bridging the Gap Between Research and Practice

RESEARCH TEAM: Geneva Tesh, 281-323-5574, genevatesh@my.UNT.edu, PhD student, Department of Learning Technologies. This study is part of a dissertation being conducted under Dr. Youngjin Lee, Professor, Department of Learning Technologies, UNT. Other committee members include Dr. Regina Kaplan-Rakowski and Dr. Bill Elieson.

The purpose of this study is to explore the use of automated essay scoring systems as a classroom tool to enhance teaching and learning ESL. The investigators will explain the study to you and will any answer any questions you might have. Taking part in this study is voluntary. The investigators will explain the study to you and will any answer any questions you might have. It is your choice whether or not you take part in this study. If you agree to participate and then choose to withdraw from the study, that is your right, and your decision will not be held against you. Your participation in this research study involves completing this brief survey about your experience using technology in the classroom, attending a 2-hour workshop via Zoom on using the software LightSide, participating in a 30-minute focus group interview immediately after the workshop, and completing another brief survey about two weeks after the workshop.

Please sign below if you are at least 18 years of age and voluntarily agree to participate in this study.

I consent to participate in this study.

I do not consent to participate in this study.

Q1 Please indicate your age.

20-29

30-39

40-49

50-59

> 60

I prefer not to answer.

Q2 Please indicate your gender.

Female

Male

Other / I prefer not to answer.

Q3 Where do you currently teach ESL?

A high school

A community college

A university

Other _____

Q4 How long have you taught ESL?

0-2 years

2-3 years

3-5 years

5-10 years

more than 10 years

Q5 Indicate how much you agree with the following statement: “My technology expertise is high.”

Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree

Q6 How do you integrate technology in your ESL writing classes? Select all that apply.

I use a learning management system such as Canvas or Blackboard.

I present lessons on an Activeboard such as White Board.

I teach remotely using a webcam and videoconferencing tool such as Zoom or Webex.

I use automated grading software.

I use plagiarism detection software.

I encourage students to use built-in support such as autocorrect.

I have students combine media forms such as digital storytelling.

I require students to collaborate on projects through cloud platforms.

I post or ask students to post to social media or blogs.

I use VR or AR technology.

Other: _____

Q7 Do you have experience do you have with machine learning, artificial intelligence, and/or learning analytics?

No

Yes (Please describe your experience.) _____

Q8 Are machine learning, artificial intelligence, and learning analytics useful for teachers?

Not at all useful

Slightly useful

Moderately useful

Very useful

Extremely useful

Q9 Do you have any experience using automated essay scoring systems in the past?

No

Yes (Please describe your experience.) _____

Q10 Is automated essay scoring useful for teachers?

Not at all useful

Slightly useful

Moderately useful

Very useful

Extremely useful

Q17 What are the pros and cons of using an automated essay scoring system?

_____

Q18 What are your expectations in terms of potential for automated essay scoring and the ways it will impact your teaching?

_____

APPENDIX B

LIGHTSIDE TRAINING MANUAL FOR ESL TEACHERS

LightSide User's Manual for ESL Teachers
Geneva Tesh
2022

PART 1. BACKGROUND INFORMATION

LightSide is a machine learning tool designed for novice users. It was developed in 2014 by Elijah Mayfied, David Adamson, and Caroline Rosé at Carnegie Mellon University.

Machine learning (ML) is the ability of a machine to imitate human behavior without being programmed. It works by taking datasets and then using algorithms and statistical models to analyze patterns in the data. The workflow follows these steps:

1. Collect, prepare, and upload data.
2. Apply an algorithm and statistical model.
3. Train the model with the data.
4. Evaluate and fine-tune the model.
5. Enter new data to predict classifications.

Automated Essay Scoring (AES) systems are machine learning programs that evaluate and score written texts. AES systems analyze essays to learn which features are related to specific scores. The system is trained to mimic human scores. The workflow follows these steps:

1. A dataset of teacher-graded essays is uploaded to the AES system.
2. The system analyzes the essays and extracts features.
3. The system develops a statistical model by working out the relationship between the features and the grades.
4. The model can be adjusted for better accuracy.
5. New essays can be uploaded, and the system will predict grades.

Teachers can build their own AES system using LightSide.

LightSide can predict essay scores, but it cannot do the following:

- replace teacher's grades (it's not accurate enough, nor is it fair to ask students to compose essays for a machine audience only)
- be used by students (too complex and it doesn't offer writing feedback beyond a single score)

Advantages of using LightSide to score students' essays:

- Immediate scores; quick identification of at-risk students
- Frees up teachers' time to focus on tasks other than scoring
- High validity and reliability, with studies showing stronger agreement between human and AES score than between two humans
- Eliminates human error factors (fatigue, distraction, bias, subjectivity, psychology, etc.)
- Open-source, completely free
- Several options for teachers to develop models based on their own grading rubrics and students' writing samples

Disadvantages:

- May increase students' testing anxiety
- May lower students' motivation when essays are graded by computer
- Essay collection/corpus creation time-consuming and difficult
- Vulnerability to cheating (system gaming)
- Teachers' learning curve in using ML platform

Machine learning is complex. Why should teachers bother learning to use such tools?

1. To provide input: ESL teachers often feel excluded when new writing tools and automated scoring systems are designed, as these tools often cater to native speaker writing. It is important for ESL teachers to actively contribute their insights to developers. Gaining a comprehensive understanding of how these tools function enables teachers to participate in the development process more effectively. ESL teachers are important stakeholders in AES and should have the opportunity to help shape and implement this technology.

2. To overcome resistance: Educational technology often goes unused due to teacher resistance, not because of problems with the technology itself. Teachers who become more familiar with new tools are more likely to embrace new technologies and use them to enhance teaching and learning.

PART 2. GETTING STARTED

Step 1. Installing Java

Before using LightSide, make sure you have Java installed on your system.

On a Mac, open the finder, then click on the following: Applications -> Utilities -> Terminal. Type “java-version” to see if Java is installed on your computer.

On Windows, open the start menu and search for “cmd.” Click the “cmd” icon, and then type “java-version” to see if Java is installed on your computer.

If you do not have Java, download it from http://java.com/download.

Step 2. Installing LightSide

Download LightSide from www.lightsidelabs.com. After downloading the correct version (Mac or Windows), open the zip file and extract it into a folder on your desktop. To open a new workspace, click on the LightSide.app on a Mac, or LightSide.bat on Windows.

Step 3. Increasing Memory

LightSide’s default memory setting is 4GB of RAM on a Mac and 1GB on Windows. The program is faster and more efficient if you change the memory settings. Follow these instructions to allocate more RAM:

On a Mac, open run.sh in a text editor. Change the value in the line MAXHEAP= “4G” from 4 to 12.

On Windows, open lightside.bat in a text editor. Change the value in the line set memory = 1G from 1 to 12.

Step 4. Formatting your data

Format your data (essays) in a single spreadsheet (Excel or Numbers, for example). The spreadsheet must contain at least two columns: a classification column (a score or grade) and a data column (essays or other written responses). You may also include columns for any other data you find useful, such as the students’ names, class, assignment details, date, etc.

Look at the following example, which uses TOEFL writing prompts:

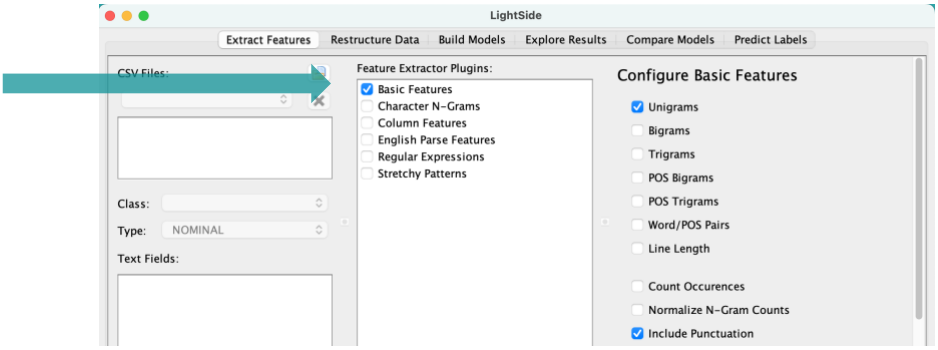
Score	Filename	Prompt	Language	Response Text
high	88.txt	P6	KOR	Some people might think that traveling in a group led by a tour guide is a good way. But, a group tour normally has its str
high	348.txt	P1	TUR	It is an important decision, how to plan your syllabus. Some students prefer to take a lot of courses and expand their kno
high	2664.txt	P2	DEU	Whether or not young people enjoy life more than older people do is an interesting question, and I for one disagree with tl

In this example, the two required columns are Score and Response Text. The score is from a human rater. Each cell under Response Text contains an entire TOEFL writing sample. The extra columns include Filename, Prompt, and Language.

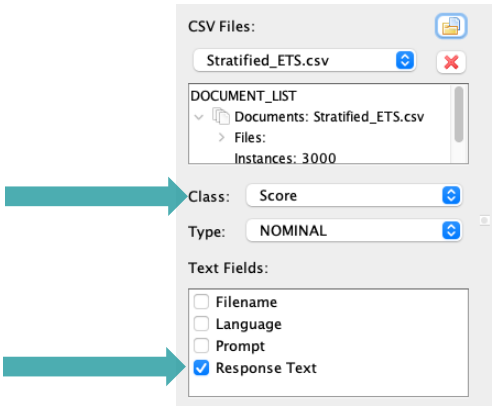
After you put all your graded essays in a single spreadsheet, convert the spreadsheet to a csv file. To do this, go to "Save as." Under File Format, choose "csv."

Step 5. Uploading data

After you format the essays in one spreadsheet and save it as a csv file, upload the data to LightSide. Open LightSide. Click the file icon to load the csv file.

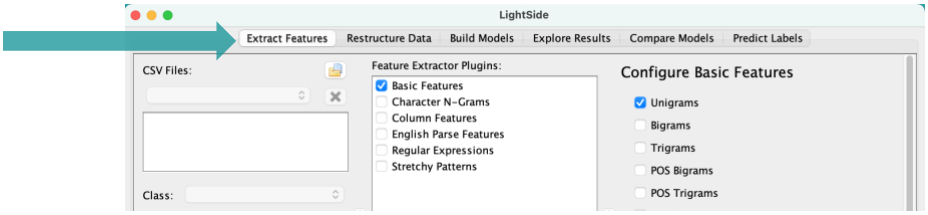


You can find information about your file here. "Instances" indicates the number of essays in your dataset (3000, in the example below). For **Class**, go to the dropdown menu and select "Score," the column containing scores. For **Text Fields**, check "Response Text."

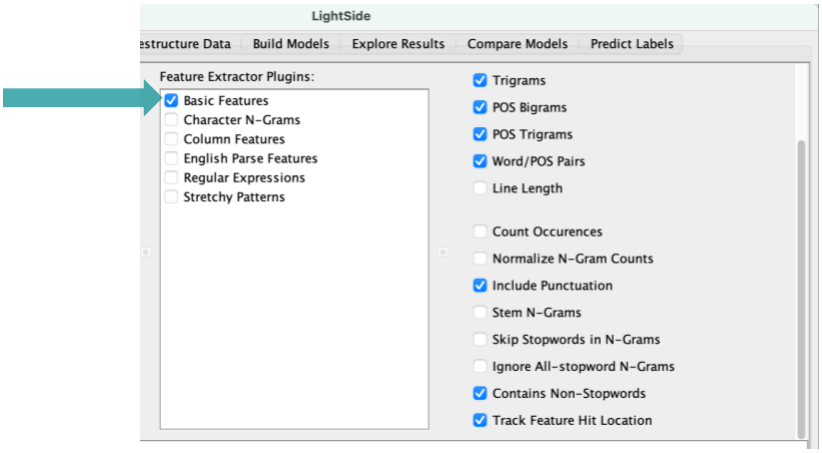


PART 3. EXTRACTING FEATURES

The first step after loading your data is extracting features. This is done on the first tab, Extract Features:



Features represent words, parts of speech, grammar, and so on. Start by clicking Basic Features from the left menu. Basic Features can extract vocabulary (n-grams), parts of speech (POS), word order, essay length, the number of times a word occurs, and punctuation.



N-grams:

In LightSide, n-grams are basically words. Checking N-grams in the Basic Features menu will indicate the presence or absence of….
Unigrams: individual words
Bigrams: 2 consecutive words
Trigrams: 3 consecutive words

Bigrams and Trigrams will catch collocations and word order.
For example, *to the mall* is not the same as *mall the to.*

POS N-grams:
In LightSide, POS refers to part of speech tags. ESL teachers are familiar with 8 parts of speech. LightSide uses computational linguistics research (Stanford POS tagger) to identify over 30 parts

of speech to distinguish different types of verbs, pronouns, etc. For example, the bigram *They talk* would be tagged as PRP_VBP, a personal pronoun followed by a non-third-person singular present verb.

POS Bigrams and POS Trigrams catch simple syntax.
Word/POS Pairs extracts a feature for every unique pairing of word and POS tag.

Other features:
Line Length counts the number of words in a document.
Count Occurrences counts the number of times a word appears in a document (by default each word gets a value of "true" if it appears at least once and "false" if it does not appear).
Normalize N-Gram Counts indicates the proportion of the document covered by a word (normalizes the occurrence of the word by length of the document).
Include Punctuation: Checks periods, commas, quotation marks, etc.

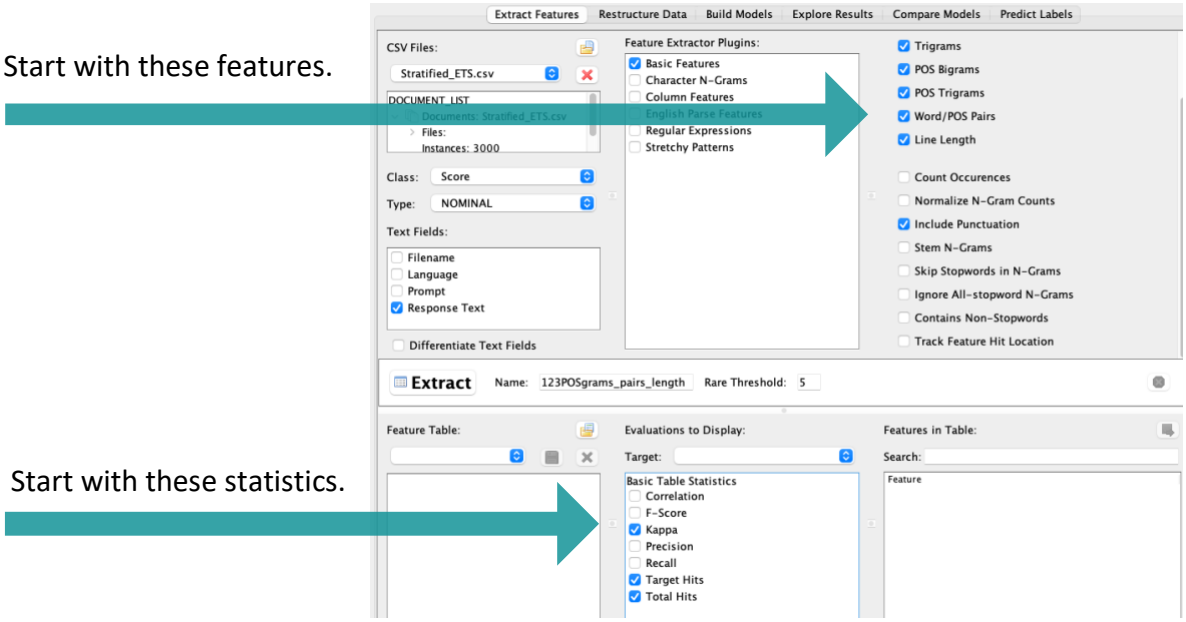
Stemming:
Stemming reduces words to a base form. For example, *informs, informed, information,* and *informant* would all be reduced to *inform.* Stemming is less extreme than lemmatization. Stemming might be useful for identifying general concepts.

Stopwords:
Stopwords are common function words that don't carry meaning. Examples include *and, the, a/an.* LightSide includes 118 stopwords.

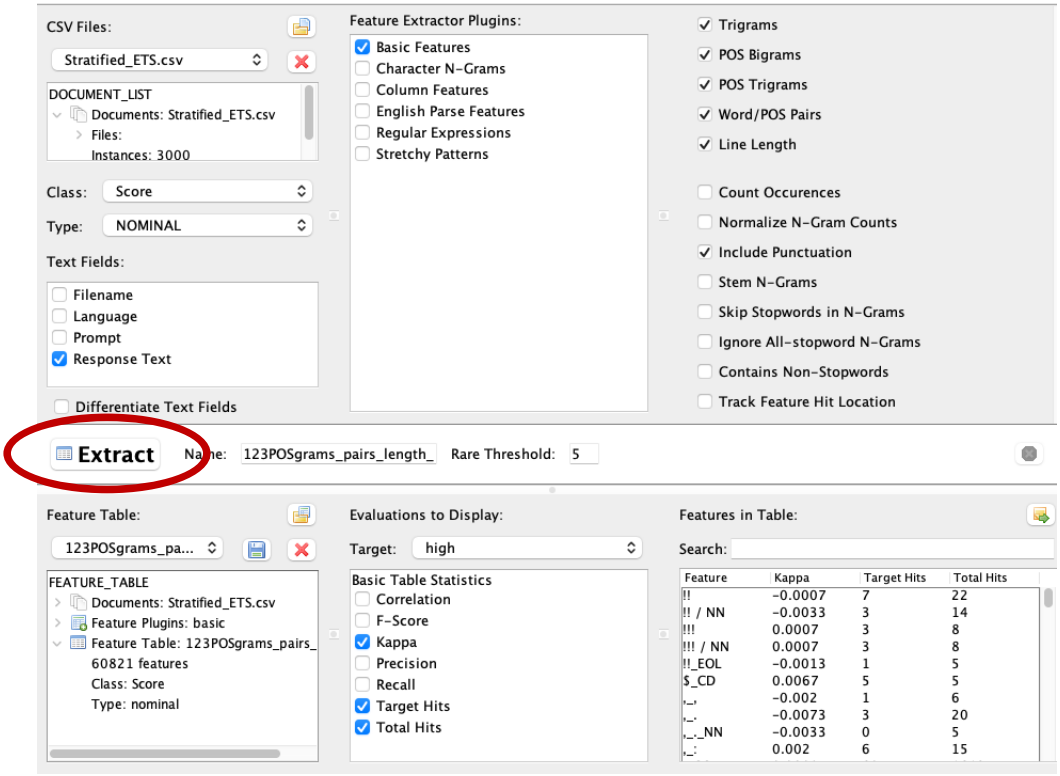
Skip Stopwords passes over stopwords; This is a good option if the task is more about content than style.
Ignore All-stopword N-Grams removes all unigram stopwords from your feature set; bigrams and Trigrams are ignored if they contain only stopwords.
Contains Non-Stopwords gives a "true" value if the essay contains at least one content word; this is not useful for scoring essays because every essay will contain at least one content word.

Select these basic features: Unigrams, Bigrams, Trigrams, POS Bigrams, POS Trigrams, Word/POS Pairs, Line Length.

In the bottom middle box, select Kappa, Target Hits, and Total Hits.
Kappa: a measure of inter-rater reliability
Target Hits: the # of times a feature appears in a class (high, medium, and low)
Total Hits: the # of times a feature appears across the entire data set



Now click on the Extract button, and wait for LightSide to complete the job. In the bottom left corner of the screen, you will find the number of features extracted in the feature table. In the example below, 60,821 features were extracted.

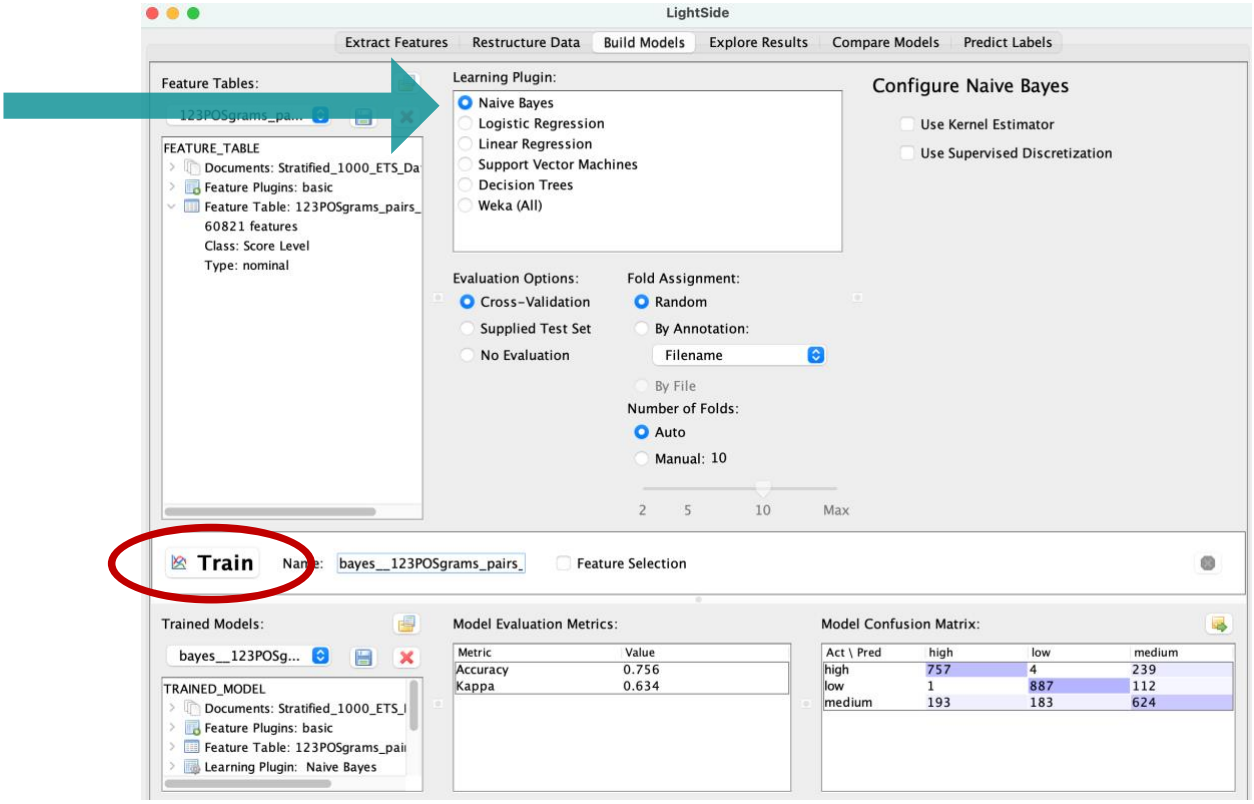


PART 4. TRAINING A MODEL

Go to the *Build Models* tab to train your model. Here, you can choose different statistical methods. For our workshop, choose Naïve Bayes (the default). The Naïve Bayes method is based on probability. If an essay has a certain set of features, what is the most likely class (score) it belongs to? Naïve Bayes will estimate the conditional probabilities for each score classification and choose the class that has the highest conditional probability for each essay.

You can also use different methods to validate your data. Start with the default option, a randomized 10-fold cross-validation.

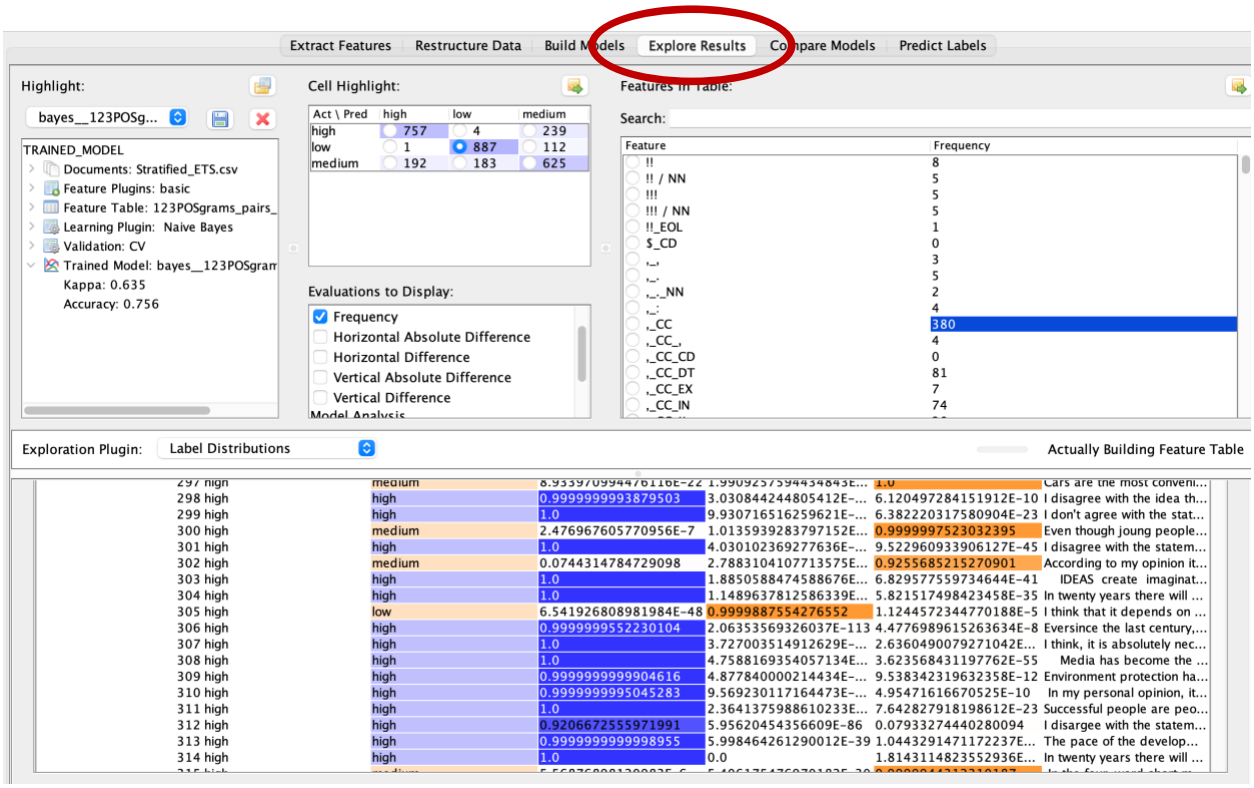
After you make your selection, click the Train button to get the results.



The Model Evaluation Metrics in the bottom middle of the screen shows accuracy results. In the example above, the model is 76% accurate with a kappa value of .63. A kappa value of .4 to .75 is considered moderate to very good (1.0 = perfect agreement). The confusion matrix on the bottom right side of the screen shows the actual scores vs. the predicted scores. In this example, we can see 757 of the 1000 high essays were accurately predicted as high. 887 low scores and 625 medium scores were predicted accurately. The greatest confusion lies in the distinction between medium and high scores.

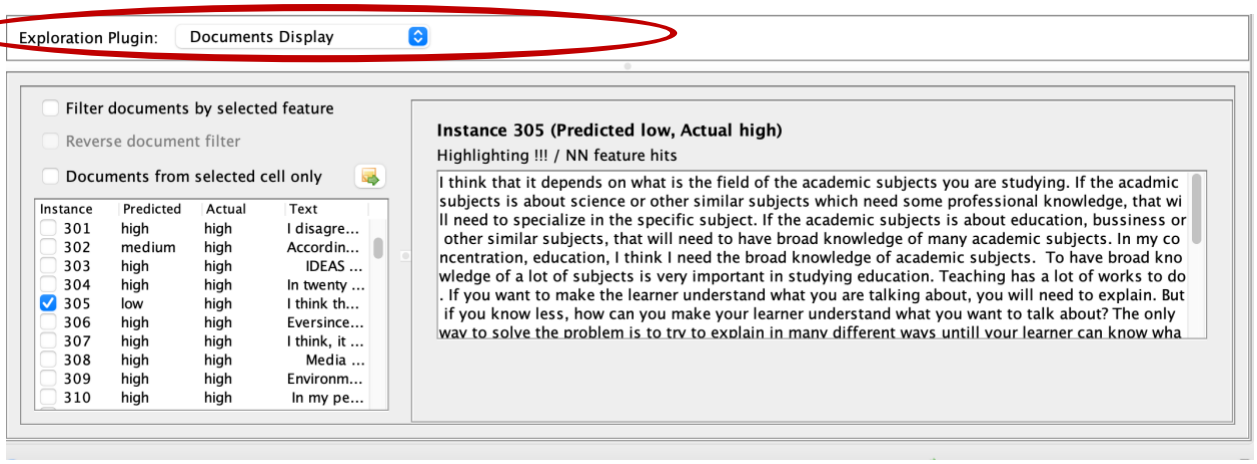
PART 5. INTERPRETING RESULTS

Use the Explore Results tab to interpret the results of your model.



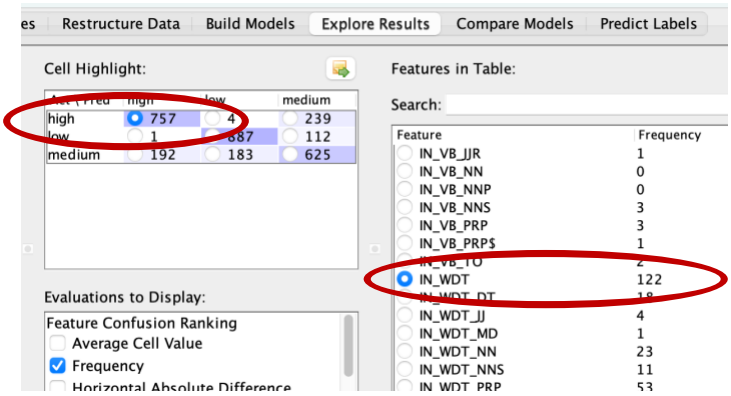
You can easily identify benchmarks and outliers in your data. The essays that are highlighted blue represent high accuracy with your model and can be used as benchmark essays to share with students or train new teachers or graders. The essays that are highlighted orange indicate weak agreement. These essays confused the model. The first column indicates the actual score, and the second column indicates the predicted score.

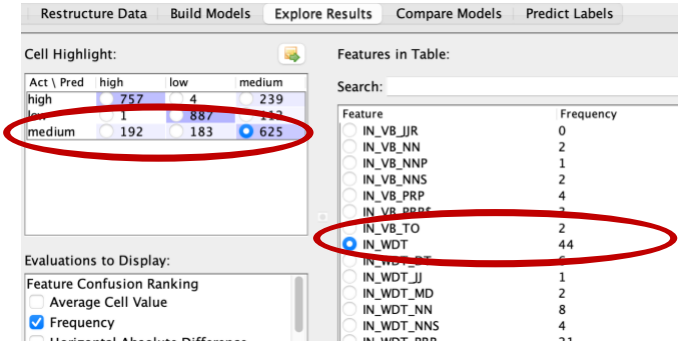
You can view entire essays by using Document Display.

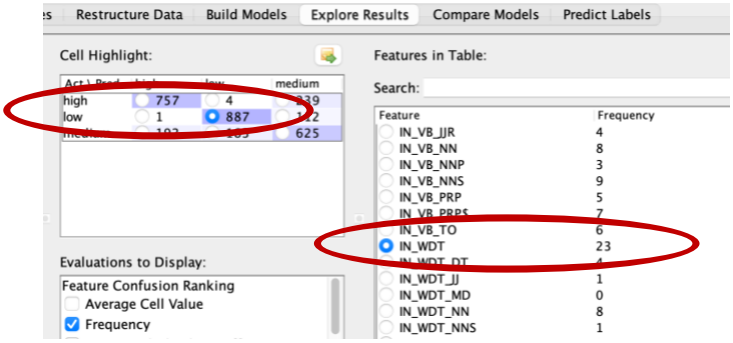


We can explore features to discover which grammatical structures and vocabulary words occur in high vs. medium vs. low scored essays. We can look at the frequency of features for different scores. Here, for example, we can see that the POS bigram IN_WDT occurs 122 times in high essays, 44 in medium essays, and only 23 in low essays.

IN = preposition; WDT = *wh*- determiner
Examples: *to whom, in which, for what*







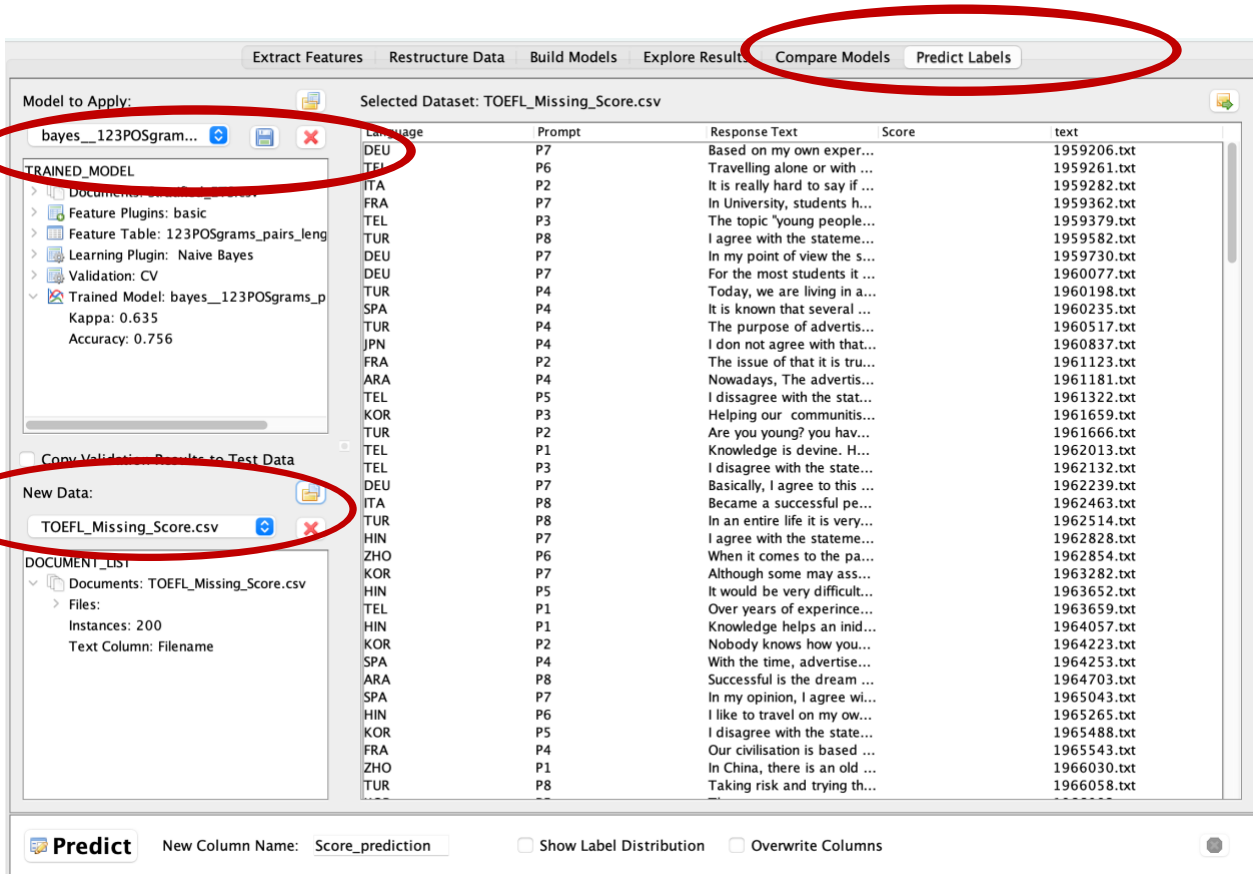
PART 6. SCORING NEW ESSAYS

To score a new set of essays, go to the Predict Labels tab.

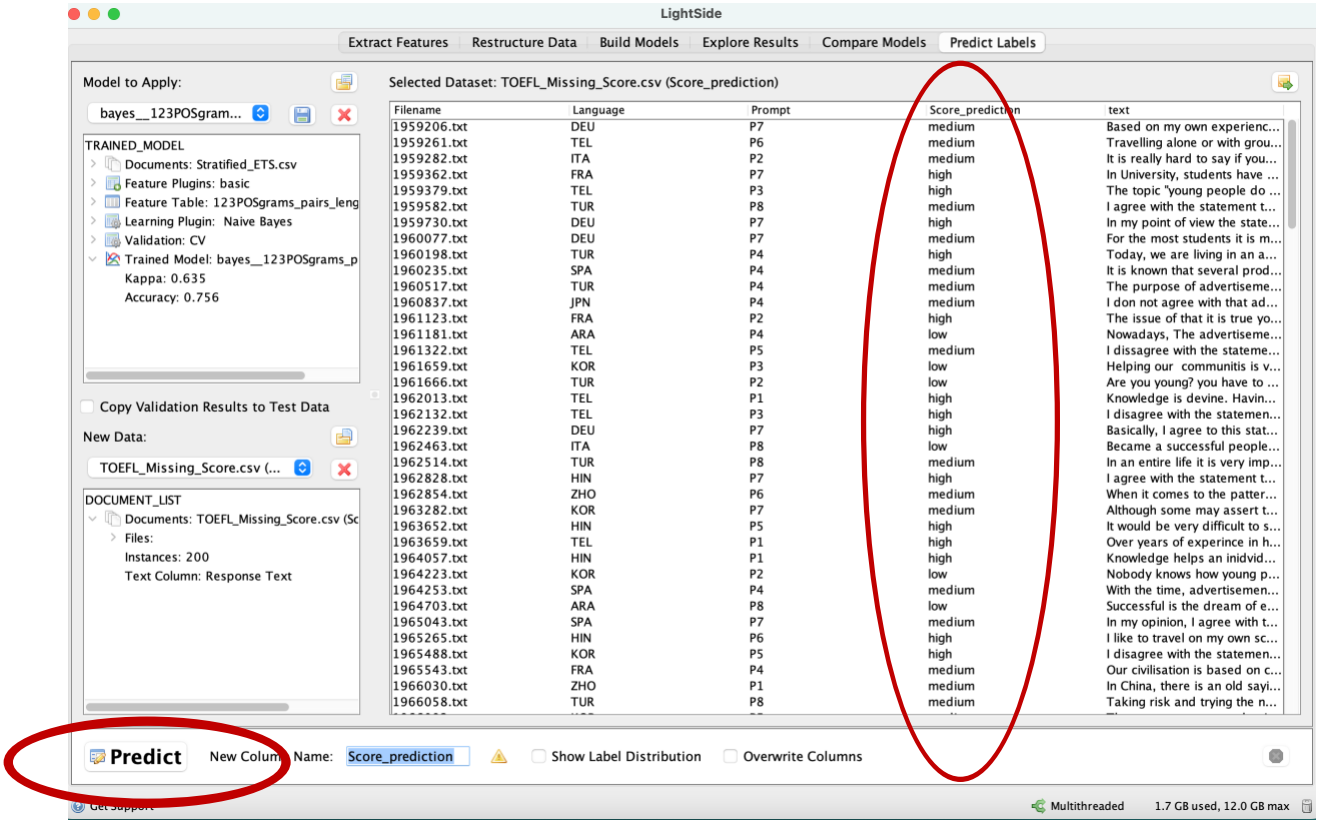
In the top left corner of the screen, choose the model you trained and built.

In the bottom left corner, load new data.

The new data needs to be in a csv file with the exact same columns as the original training set, but *without* scores. The new data needs to be in a csv file with the exact same columns as the original training set, but without scores.



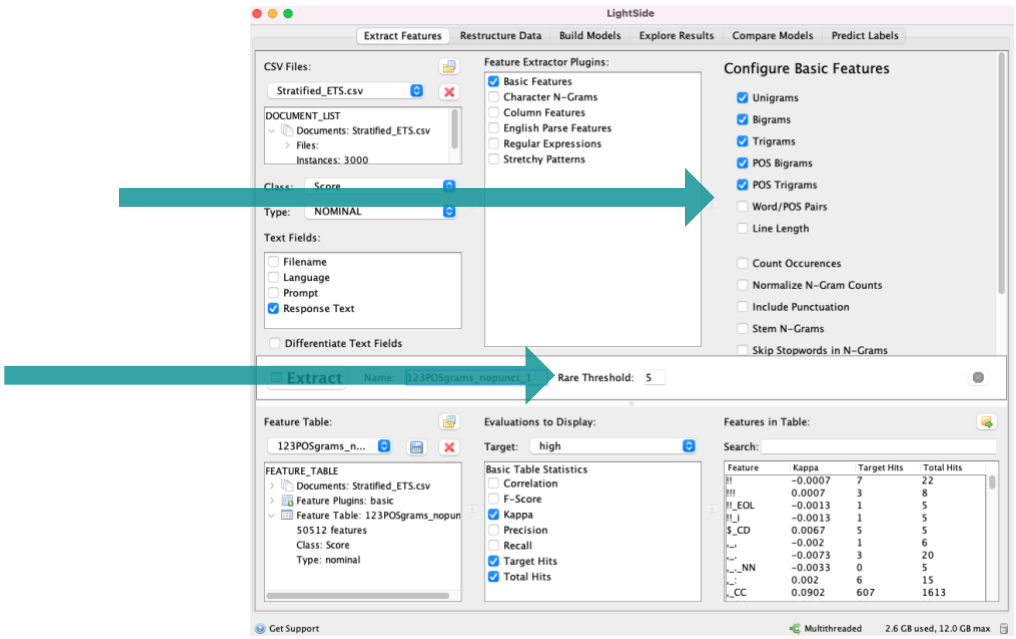
Click the **PREDICT** button, and now you will get a new column with the machine predicted scores.



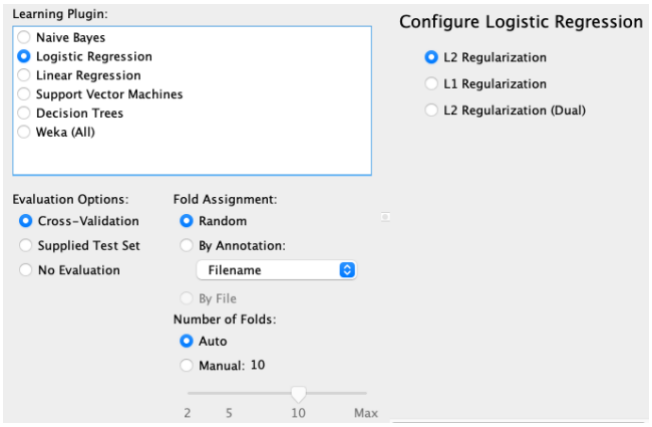
PART 7. IMPROVING THE MODEL

There are lots of ways to optimize your model through the "Extract Features" and "Build Models" tabs. You can reduce the number of features you choose to extract. This will yield a table with fewer features. This can sometimes be helpful in reducing "noise."

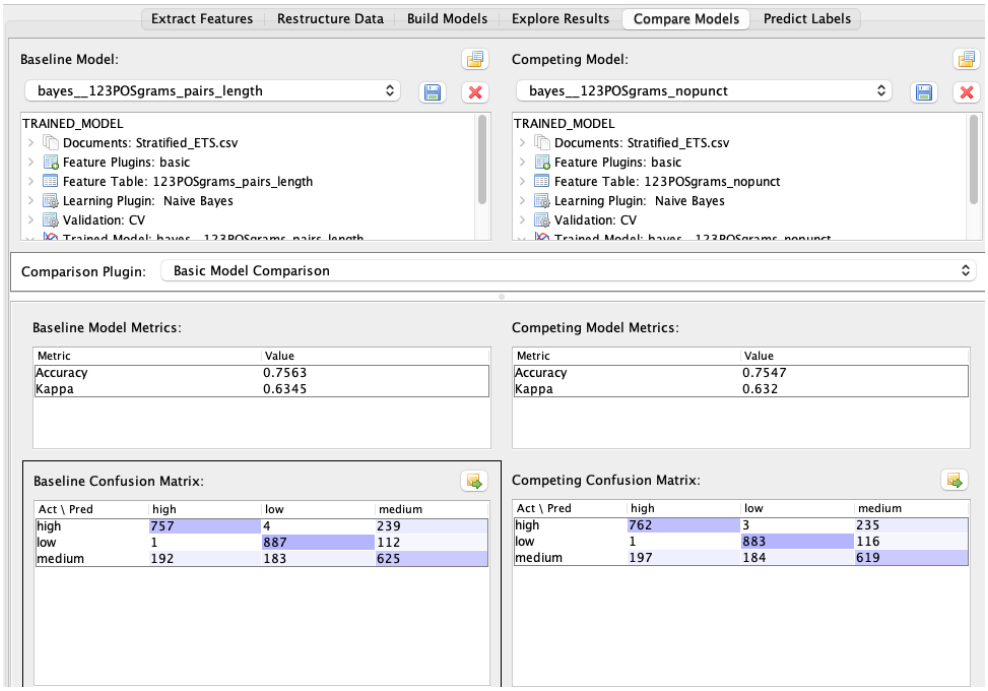
Another option is changing the "Rare Threshold," which indicates the number of times a feature appears across the dataset. The default is 5, meaning the feature needs to appear only 5 times across the entire set of essays.



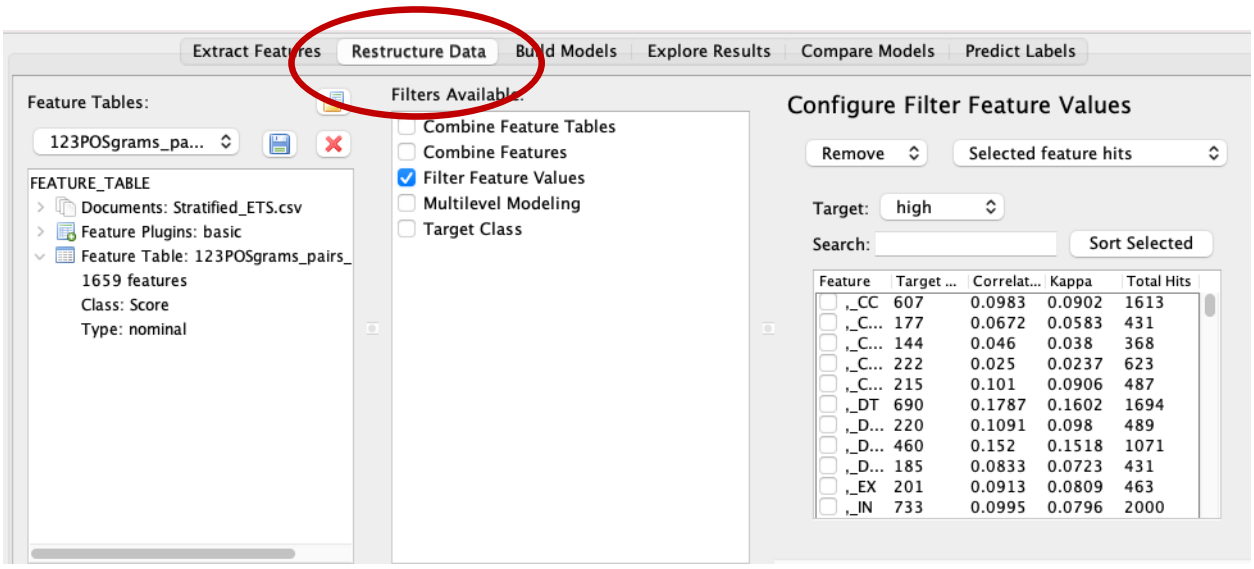
You can also try different statistical models, such as Logistic Regression or Support Vector Machines, on the Build tab.



As you try different features and statistical models, you can easily compare models with the "Compare Models" tab. This shows two models side by side. You can pick any saved models from the dropdown menu.



The "Restructure Data" tab allows even more advanced optimization. For example, you can combine features or filter out specific features.



LightSide provides endless opportunities for optimizing your own AES model. Have fun exploring!

REFERENCES

Alamari, B. (2021). Challenges of implementing technology in ESL writing classrooms: A case study. *English Language Teaching 14*(12), 36-43.

Almarzooq, Z., Lopes, M., & Kochar, A. (2020). Virtual learning during the COVID-19 pandemic: A disruptive technology in graduate medical education. *Journal of the American College of Technology, 47*(4), 778-786.

Al-Wasy, B.Q. (2020). The effectiveness of integrating technology in EFL/ESL writing: A meta-analysis. *Interactive Technology and Smart Education 17*(4), 435-454.

Amaral, L. & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL 23,* 4-24.

Anson, C. (2006). Can't touch this: Reflections on the servitude of computers as readers. In Ericsson, P.F. & Haswell, R. (eds.) *Machine scoring of human essays*, 38-56. Logan, UT: Utah State University Press.

Antoniadis, G. & Desmet, P. (2016). NLP for learning and teaching: challenges and opportunities. Revue TAL, Association pour le Traitment Automatique des Langues.

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D. (2015). Look together: Analyzing gaze coordination with epistemic network analysis. Frontiers in Psychology, 6(1016).

Arastoopour, G., Swiecki, Z., Chesler, N. C., & Shaffer, D. W. (2015). Epistemic Network Analysis as a tool for engineering design assessment. Presented at the American Society for Engineering Education, Seattle, WA.

Arastoopour, G., Shaffer, D. W., Swiecki, Z., Ruis, A. R., & Chesler, N. C. (2016). Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *International Journal of Engineering Education*, 32(3B), 1492–1501.

Bengio, Y. (2002). *New distributed probabilistic language models*. Université de Montréal.

Bhandari, P. (2022). Construct validity: Definitions, types, and examples. *Scribbr.*

Blood, I. (2011). Automated essay scoring: A literature review. Apple Award Winning Papers. In *TESOL & AL*. *11*(2).

Charya, N.; Doshi K.; Bawkar, S; Shankarmani, R. (2015). Intrinsic plagiarism detection in digital data. *International Journal of Innovative and Emerging Research in Engineering*, *2*(3), 23-30.

Chen, M.H., Chen, W.F., & Ku, L.W. (2018). Application of sentiment analysis to language learning. *IEEE Access,* 6.

Chen, J., Zhang, M., & Bejar, I.I. (2017). An investigation of the e-rater scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model (Research Report No. RR 17-04). Princeton, NJ: Educational Testing Service.

Cho, H. (2014). What writing tasks do TESOL professors require? *TESOL Journal*, 247–264.

Christensen, L. (2003). The politics of correction: How we can nurture students in their writing and help them learn the language of power." *The Quarterly 25*(4), 6-9.

Crossley, S. A., Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining*, *8*(2), 1-19.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42, 475–493.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment 5*.

Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal, 28*(1), 99-134.

DiSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), Constructivism in the computer age (pp. 47–70). Hillsdale, NJ: Erlbaum.

Eagan, B., & Hamilton, E. (2018). Epistemic Network Analysis of an International Digital Makerspace in Africa, Europe, and the US. Presented at the Annual Meeting of the American Education Research Association, New York, NY.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Fillmore, C. (1968). The case for case. In Bach and Harms (eds.) *Universals in Linguistic Theory.* New York: Holt, Rinehart, and Winston.

Fleckenstein, J., Leucht, M., Pant, H.S., & Koller, O. (2016). Proficient beyond borders: Assessing non-native speakers in a speakers' framework. *Large-scale Assessments in Education, 4*(19).

Fu, S., Gu, H., & Yang, B. (2020). The affordances of AI-enabled automatic scoring applications on learners' continuous learning intentions: An empirical study in China. *British Journal of Educational Technology, 51*(5), 1674-1692.

Gartner Inc. (2021). The Hype Cycle: Understanding the pitfalls and opportunities of innovations. Gartner Research.

Goldberg, G.I. (2012). Judgement-based scoring by teachers as professional development: Distinguishing promises from proof. *Educational Measurement: Issues and Practice, 31,* 38-47.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In Sweet, A.P. & Snow, C.E. (eds.), *Rethinking reading comprehension*. New York: Guilford Publications.

Green, A. (2020). Washback in language assessment. In Chapelle, C. (ed.) *The encyclopedia of applied linguistics.* Wiley Blackwell.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 4-43.

Ha, M., Nehm, R.H., Uraban-Lurain, M., & Merrill, J.E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE—Life Sciences Education 10*, 379-393.

Halliday, M.A.K. (1973). *Explorations in the Functions of Language.* London: Edward Arnold.

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications*, 15(5), 22-37.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Higgins, D., Burstein, J., and Attali, Y. (2006). Identifying off-topic students essays without topic-specific training data. *Natural Language Engineering, 12*(2), 145–159.

Hyland, K., & Hyland, F. (2019). *Feedback in Second Language Writing: Contexts and Issues*. Cambridge University Press.

Hymes, D. (1971). Competence and performance in linguistic theory. In Huxley, R. & Ingram, E. (eds.) *Language Acquistion: Models and Methods,* 3-28. Academic Press.

Johnson, W. L. (2007). Serious use of a serious game for language learning. In *Proceedings of AIED.*

Jones, K.S. (1994). Natural language processing: A historical review. In Zampolli, A., Calzolari, N., & Palmer, M. (eds) *Current Issues in Computational Linguistics: In Honour of Don Walker*, Linguistica Computazionale, vol. 9, Springer, Dordrecht.

Jurafsky, D. & Martin, J.H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice Hall, 2nd ed.

Karim, K., & Nassaji, H. (2020). The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research, 24*(4), 519–539.

Kay, M. (1979). Functional grammar. In *Proceedings of the Berkeley Linguistics Society.* Linguistics Society of America.

Klebanov, B.B. & Madnani, N. (2022). *Automated essay scoring*. Springer Nature Switzerland.

Kolowich, S. (2014). Writing instructor, skeptical of automated grading, pits machine vs. machine." *The Chronicle of Higher Education*. *LX*(33), A12.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education.*

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum.

Leckie, G. & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity, drift, central tendency, and rater experience. *Journal of Educational Measurement, 48,* 399-418.

Lee, L. (2003). "I'm sorry, Dave, I'm afraid I can't do that": Linguistics, statistics, and natural language processing circa 2001. *National Research Council on the Fundamentals of Computer Science.*

Leki, I. (1990). Potential problems with peer responding in ESL writing classes. *CATESOL Journal,* 5-19.

Liang, M., & Guo, Y. (2020). Automated essay scoring: Applications to educational technology. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Modern Educational Technologies, Applications, and Management* (pp. 1-27). IGI Global.

Liddy, E. (2001). Natural language processing. In *Encyclopedia of Library and Information Science*, 2nd ed. Marcel Decker.

Link, S., Dursun, A., Karakaya, K., Hegelheimer, V. (2014). Towards best practices for implementing automated writing evaluation. *CALICO Journal 31:* 3, 323-344.

Linn, M. C., Eylon, B.-S., & Davis, E. A. (2004). The knowledge integration perspective on learning. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 29–46). Mahwah, NJ: Lawrence Erlbaum Associates.

Litman, D. (2016). Natural language processing for enhancing teaching and learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI-16).

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EduCause Review, 46*:5, 31-40.

Lu, X. & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29,* 16-27.

Lynch, M. (2019). Using feedback loops to impact student learning. *The Tech Edvocate*.

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. W. (2018). Epistemic Network Analysis (Version 1.7.0) [Software]. Available from http://app.epistemicnetwork.org

McNamara, D., Crossley, S., & Roscoe, R. (2012). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428- 012-0258-1

Meurers, D. (2019). Natural language processing and language learning. In Chapelle, C.A. (ed.) *Concise encyclopedia of applied linguistics.* Wiley.

Osborne, D. (2015). The ugly stepchild: On the position of ESL programs in the academy. *College ESL Quarterly*. Language Arts Press.

Pennington, M. (2011). The impact of the computer in second-language writing. *Second-language writing in the composition classroom: A critical sourcebook*. Boston: Bedford/St. Martins, 2011.

Perelman, L. (2013). Critique of Mark D. Mark D. Shermis & Ben Hamner, contrasting state-of-the-art automated scoring of essays: Analysis." *The Journal of Writing Assessment*.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skills. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford, England: Blackwell.

Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse: A computer corpus-based comparison of non-native and native English. In *Corpus-based and computational approaches to discourse anaphora*, eds. Botley, S. & McEnery, T.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, *Human Language Technologies, 1*.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kulkich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Behavior, 18,* 103-134.

Quillian, M. (1968). Semantic networks. In *Semantic information processing,* MIT Press.

Racynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring: Which essays? Which human raters? Which scores? *Applied Measurement in Education, 31*(3), 233-240.

Racynski, K., Cohen, A., Engelhard, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame of reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement, 52,* 301-318.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation, *Educational Psychology*, *37*:1, 8-25.

Ruis, A. R. (2018). "Trois Empreintes d'un Même Cachet": Toward a Historical Definition of Nutrition. In E. T. Ewing & K. Randall (eds.), Viral networks: Connecting digital humanities and medical history (pp. 179–212). Blacksburg: VT Publishing.

Ruis, A. R., Rosser, A. A., Quandt-Walle, C., Nathwani, J. N., Shaffer, D. W., & Pugh, C. M. (2018). The hands and head of a surgeon: Modeling operative competency with multimodal epistemic network analysis. American Journal of Surgery, 216(5), 835-840.

Schank, R. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology, 3*(4), 552-631.

Shaffer, D. W. (2004). Pedagogical praxis: Using technology to build professional communities of practice. Association for Computing Machinery (ACM) SigGROUP Bulletin, 24(3), 39–43.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers and Education*, 46(3), 223–234.

Shaffer, D. W. (2007). How computer games help children learn. New York, NY: Palgrave Macmillan.

Shaffer, D. W. (2012). Models of situated action: Computer games and the problem of transfer. In C. Steinkuehler, K. D. Squire, & S. A. Barab (Eds.), *Games, learning, and society: Learning and meaning in the digital age* (pp. 403–431). Cambridge, UK: Cambridge University Press.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3), 9–45.

Shaffer, D. W., Hatfield, D. L., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E. A., … Frank, K. (2009). Epistemic network analysis: A prototype for 21st century assessment of learning. *International Journal of Learning and Media*, 1(1), 1–21.

Shaffer, D. W., & Ruis, A. R. (2017). Epistemic network analysis: A worked example of theory-based learning analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gasevic (Eds.), Handbook of learning analytics (pp. 175–187). Society for Learning Analytics Research.

Shankar, R. S., & Ravibabu, D. (2018). Digital report grading using NLP feature selection. Soft computing in data analytics. *Advances in intelligent systems and computing*, 615-623.

Shermis, M.D. (2014). State of the art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76.

Shermis, M.D., & Burstein, J. (2013). *Handbook of automated essay evaluation: current applications and new directions*. Routledge.

Shermis, M.D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays: Analysis. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). Routledge.

Siebert-Evenstone, A., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A. R., & Williamson Shaffer, D. (2017). In Search of Conversational Grain Size: Modelling Semantic Structure Using Moving Stanza Windows. *Journal of Learning Analytics*, 4(3), 123–139. https://doi.org/10.18608/jla.2017.43.7

Snow, M. (1991). Content-based instruction: A method with many faces. In Alatis, J. (ed.) *Linguistics and language pedagogy,* 461-70. Georgetown University Press.

Warner, J. (2018) *Why they can't write: Killing the five-paragraph essay and other necessities*. Johns Hopkins University Press.

Weigle, S. (2013). English as a second language writing and automated essay evaluation. In Shermis, M.D., & Burstein, J. (eds.) *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Wenger, E. (1999). Communities of practice: Learning, meaning, and identity. Cambridge, UK: Cambridge University Press. Retrieved from http://books.google.com/books?id=heBZpgYUKdAC&pgis=1

Wiggins, G. (2012). Seven keys to effective feedback. *ASCD, 70*:1.

Wilkins, D. (1976). *National syllabuses.* Oxford University Press.

Wilks, Y. (1973). Preference semantics. Advanced Research Projects Agency, National Technical Information Service.

Wooldridge, A. R., Carayon, P., Eagan, B. R., & Shaffer, D. W. (2018). Quantifying the qualitative with epistemic network analysis: A human factors case study of task-allocation communication in a primary care team. IIE Transactions on Healthcare Systems Engineering.

Yang, Y., Buckendahl, C.W., Juszkiewicz, P.J., & Bhola, D.S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*, 391-412.

Yannakoudakis, H., Briscoe, T., and Medlock, B. 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*.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly,* 19(1), 79-101.

Zheng, Y., & Cheng, L. (2008). College test in China. *Language Testing, 25*, 408-417.

Zhu, M., Liu, O.L., & Lee, H.S. (2019). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Educational Testing Services: The Concord Consortium.

Zupanc, K., & Bosnić, Z. (2018). Increasing accuracy of automated essay grading by grouping similar graders. *Proceedings of the 8th International Conference*, *Web Intelligence, Mining and Semantics, 18*.