2018-05-01

Applying the Developmental Path of English Negation to the Automated Scoring of Learner Essays

Allen Travis Moore
Brigham Young University

Applying the Developmental Path of English Negation to the Automated Scoring of Learner Essays

Allen Travis Moore

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Deryle Lonsdale, Chair
Troy L. Cox
Robert Reynolds

Department of Linguistics and English Language

Brigham Young University

ABSTRACT

Applying the Developmental Path of English Negation to the Automated Scoring of Learner Essays

Allen Travis Moore
Department of Linguistics and English Language, BYU
Master of Arts

The resources required to have humans score extended written response items in English language learner (ELL) contexts has caused automated essay scoring (AES) to emerge as a desired alternative. However, these systems often rely heavily on indirect proxies of writing quality such as word, sentence, and essay lengths because of their strong correlation to scores (Vajjala, 2017). This has led to concern about the validity of the features used to establish the predictive accuracy of AES systems (Attali, 2007; Weigle, 2013). Reliance on construct-irrelevant features in ELL contexts also forfeits the opportunity to provide meaningful diagnostic feedback to test-takers or provide the second language acquisition (SLA) field with real insights (C.-F. E. Chen & Cheng, 2008). This thesis seeks to improve the validity and reliability of an AES system developed for ELL essays by employing a new set of features based on the acquisition order of English negation. Modest improvements were made to a baseline AES system's accuracy, showing the possibility and importance of engineering features relevant to the construct being assessed in ELL essays. In addition to these findings, a novel ordering of the sequence of English negation acquisition not previously described in SLA research emerged.

Keywords: automated essay scoring, English negation developmental sequence, AES validity, order of acquisition

ACKNOWLEDGEMENTS

Without question, my biggest debt of gratitude goes to my wife, Sydney, who for the better part of two years has supported me in pursuing this master's degree and in writing this thesis. She has, with patience and understanding, spent many evenings alone putting three rambunctious children to bed giving me the chance to be as focused and as dedicated as I needed to be to accomplish these tasks. I will always look up to her example of selfless service and sacrifice.

Dr. Deryle Lonsdale also took me under his wing and mentored me through this process in ways that were both outside of his job description and above and beyond my expectations. The expectations he had of me caused me to stretch and grow in ways I never thought I could. Not only is he a great teacher, but I count him as a true friend.

I am indebted to Dr. Troy Cox for connecting me to this project and giving me the time and resources I needed to develop the automated essay scoring system featured in this thesis. His expertise in the field of language assessment also pushed me to be more exhaustive in my research and more precise in my writing.

Dr. Robert Reynolds was someone I relied on heavily during the development phase of the automated essay scoring system. Frequently, I would work on a programming problem for days only to finally go to Dr. Reynolds for guidance where I would be enabled to solve the issue in mere minutes. His careful explanations demystified concepts as complex as machine learning to someone who had only taken a single programming class at the time.

I acknowledge that I would not have made it to this point without any one of these remarkable individuals. Their contribution and service have been crucial to my success.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

The automated scoring of essays has been a research topic for some time in computational fields in education and industry. Increased accuracy in automated scoring is important both to reduce costs associated with expert grading and to ensure consistency in assessing writing quality, content, and/or language proficiency (Dikli, 2006). The implications, therefore, have proven to be quite far-reaching.

The process of developing an automated essay scoring (AES) system requires a set of expert graded essays from which a set of features that are deemed salient to the scores are extracted. These features are then statistically analyzed relative to the essay scores (most often through computational means) and a scoring model is built, which can be used to assign a grade based on the realization of these features in new essays. A successful AES system is able to produce scores that are reasonably close to human ratings with a high level of consistency. An example is Educational Testing Service's *e-rater* system^{TM} (hereafter referred to as *e-rater*), which produces scores that either match or are within one point of human scores on a six point scale about 96% of the time (Chodorow & Burstein, 2004).

While demand for such systems has grown in an industry where educators have more and more responsibilities placed on them, concerns have balanced the breadth of their implementation. Public concerns tend to relate to a more general skepticism of technology being awarded tasks previously only held by humans (see Attali (2013) for a full account of this type of critique). In a 1999 New York Times article, Janny Scott famously quipped, "It has come to this. The essay, the great literary art form that Montaigne conceived and Virginia Woolf carried on […] has sunk to a state where someone thinks it is a bright idea to ask a computer if an essay is any good" (Scott, 1999). As Briscoe, Medlock, & Andersen (2010) relate, "this type of philosophical objection tends to dissipate as algorithms become more effective at any given task" (pg. 11).

From the perspective of researchers and test administrators, questions have been raised concerning the validity of the features that determine the predictive accuracy of AES systems. For example, while a feature that counts the number of words in the essay is used in almost all AES systems because of its strong correlation with human ratings, generic length measures show little about the linguistic aptitude of the learner (Attali, 2007). Furthermore, it is uncommon for scoring rubrics that expert raters use to even mention essay length as a criterion of evaluation (Chodorow & Burstein, 2004). Yang, Lu, & Weigle (2015) have voiced particular concern with reliance on length measures when scoring essays written by ESL students as they show little in areas like syntactic complexity, which should be of greater interest in these contexts. The evidence of incongruence between some of the commonly used features in AES systems and the constructs actually being assessed has caused the research community to give preeminence to those features that have some connection to the purpose of the written assessment (Weigle, 2013). AES researchers and developers must continue to take steps to address these concerns with empiricism and make changes where necessary.

My thesis addresses this problem in the context of the automated scoring of English language learner (ELL) essays. The goal is to determine how features related to the developmental patterns of English acquisition can be used to increase the performance of an AES system built for ELL essays. Specifically, the developmental path of English negation is the focus of this thesis and is used to address the following research questions.

RQ1: Do the developmental patterns of the acquisition of negation surface in ELL essays? If so, how?

RQ2: What is the effect of including features based on negation acquisition patterns in a fully functional AES system?

The purpose of written examinations in the context of second language learning is usually to directly assess the language proficiency of the learner and less so their rhetorical style or the essay's content (Weigle, 2013). For this reason, I propose that features related to an area of acquisition for which there are already rich qualitative descriptions in second language acquisition (SLA) literature would help to increase the validity of AES for ELL essays.

Negation has been used as a pointer to linguistic phenomena in multiple studies (see e.g., de Swart, 2009). It has also been shown to have a "systemic, non-linear and unevenly paced development along predictable stages" (Ortega, 2014, pg. 119). Besides negation, several other systematic patterns, including the acquisitional development of English question forms, auxiliaries, and the verb system, have been uncovered and described in a similar fashion (see Dulay & Burt, 1974; Krashen, 1985; Pienemann, Johnston, & Brindley, 1988). If the path of English negation acquisition proves salient to the prediction of scores for ELL essays, a complete set of features based on second language (L2) English development patterns could ostensibly be built to improve the accuracy and help reinforce the validity of AES in these contexts.

Finding these construct-relevant features by which to train AES systems is important in order to show that "automated scores do not amount to counting words" (Attali, 2007, pg. 2). These features could provide this and other automated scorers built for ELL essays in the future with an external point of reference directly related to language acquisition. As Burt & Dulay

(1978) have noted, this would "[permit] actual levels of language development to be incorporated to the proficiency levels" (pg. 183). And, as providing feedback has become important in this field, these kinds of features could be used to give students and instructors relevant insight into the state of their proficiency.

Chapter 2 - Literature Review

This chapter is divided into a review of the literature related to developmental paths of English acquisition followed by a comprehensive overview of the state of AES. I begin by outlining the concept of interlanguage and how it grew out of the idea that language learner errors are not completely random. I then discuss how this has led to research that attempts to uncover whether there is an order to the acquisition of certain L1 and L2 structures. Next, to further elucidate the methodological practices in this area, I review several attested L2 acquisition orders of English structures. Finally, I introduce the common practices, procedures and concerns related to the use of AES in both L1 and L2 contexts.

2.1 Interlanguage Theory

Some central tenets in current SLA theory are that language acquisition is systematic and learner errors are not necessarily random (White, 2008; Ortega, 2014). These ideas stemmed from the work of researchers like Selinker (1972) and his contemporaries who were the first to describe what is now widely known as ‘interlanguage’, which Selinker defined as, “a separate linguistic system based on the observable output which results from a learner’s attempted production of a [target-language] norm” (pg. 214). He developed this theory based on the fact that second language learners’ attempted productions in their target language often result in structures that are found neither in their native language nor in the target language. This theory formed the hypothesis that interlanguage could be studied on the same grounds as a natural language. If interlanguage is itself a grammar with systematic rules, questions naturally arise concerning its structure, development, and constraints. Hawkins (2008) outlined two foundational questions that define research in this area: (1) what are the properties of an L2 grammar that give rise to observed performance, and (2) how does an L2 learner arrive at that grammar?

The two main theories that attempt to tackle these question are the nativist and emergentist approaches. The nativist approach claims that observed L2 performance is not merely a result of the learner's individual experience and knowledge in their L1, but that humans have an innate linguistic apparatus into which all natural languages fit and by which they are constrained. This approach has strong ties to the theory of universal grammar (UG). Therefore, if interlanguage is itself a linguistic system, its structure should emerge and be constrained in the same way as all natural languages. Studies have shown that interlanguage performances often prove viable in other natural languages even when the observed performance is not found in either the L1 of the speaker or the targeted L2 grammar (Hawkins, 2008). This, they claim, proves the existence of the innate system.

Emergentists believe that there is not a separate linguistic system that structures and constrains natural languages, but that language emerges from existing processes in the human mind (Ellis, 2008). The emergentist thesis for language states: "The phenomena of language are best explained by reference to more basic non-linguistic (i.e., 'non-grammatical') factors and their interaction—physiology, perception, processing, working memory, pragmatics, social interaction, properties of the input, the learning mechanisms, and so on" (O'Grady, 2008, pg. 448). Learning mechanisms include statistically driven tendencies to acquire language based on transitional probabilities (the predictability of adjacent forms), salience, and analogical generalizations among others (Casillas, 2008; Hawkins, 2008). O'Grady (2008) emphasizes that these processes do, in fact, collectively contribute to the species-specific ability to acquire language, but that acquiring language is not their only role in the human system. In other words, these distinct abilities come together based on what could be considered an innate human desire or need to communicate.

2.2 Acquisition Order

While a unification of the above theories dealing with the origin and structure of the interlanguage system has yet to come about, researchers have attempted to investigate the acquisition order of certain linguistic structures by analyzing the interlanguage of learners. Analyses of this sort can be performed with longitudinal or cross-sectional data made up of participant responses to some language elicitation task. Interestingly, a variety of studies have shown that an attested acquisition order of a particular structure will be mostly identical across spoken and written domains (Dulay, Burt, & Krashen, 1982). There are several best practices to follow when attempting uncover an acquisition order, which will be discussed in this section.

Table 1 shows an example of what is generally understood about the L2 acquisition order of English grammatical morphemes. In general, the present progressive marker *-ing* gets acquired first followed by the plural *-s*, and so on. However, Krashen (1985) reviewed a number of studies that claimed to find no common order related to the acquisition of these morphemes. In investigating why, he hypothesized that the more a research task constrained a learner's speech or writing, the less natural their productions would be, essentially masking the acquisition order. Further analysis indeed revealed that discrete-point tasks such as multiple choice questions and completion items had the effect of bringing the learner's "conscious mind" into the foreground, overriding their natural communicative abilities. Dulay et al. (1982) noted that these types of methodological errors stem from a "failure to distinguish between purposes of linguistic manipulation tasks and natural communication tasks" (pg. 225). Importantly for the purposes of this thesis, timed writing tasks fall into the category of natural communication and therefore do not disturb the emergence of an acquisition order, but in fact have been shown to confirm them (Krashen, 1985).

Stage	Morpheme
1	*-ing* Plural *-s* *Be* copula
2	*Be* auxiliary *a/the*
3	Irregular past
4	Regular past *-ed* Third person *-s* Possessive *-'s*

Table 1: ELL morpheme acquisition order adapted from Ortega (2014, pg. 125)

Structures under investigation like those in Table 1 can be rank ordered relative to one another by calculating a test group's percentage of accuracy with each one individually. This is done by calculating all obligatory occasions for a given structure in the corpus over the participants' actual use of the structure in those contexts. A potential pitfall when it comes to describing an acquisition sequence is in rank ordering closely related structures that may actually be different manifestations of the same stage of development or might be ordered differently with more representative data (Dulay et al., 1982). Some of the methods used to combine or disambiguate these types of structures are the hierarchical analysis and the use of confidence intervals.

Acquisition as defined by Dulay & Burt (1974) is the correct use of a given structure in 90% of obligatory cases. Hierarchical analysis pairs each structure under investigation together (*a, b*) and gets the percentage of time that the first structure is acquired by the learner when its pair is not. If structure *b* is acquired before *a* in ≤5% of cases, they are considered separate and follow the sequence *a, b*. Confidence intervals can show where structures will fall relative to one another based on the mean of their ranking or some other metric for a group of participants. These intervals show where the mean is with 90, 95, or 99% confidence.

A final pitfall for studies looking into acquisition orders is the occasional sparse representation of some structures in the data. Historically, the methods used to uncover and describe these phenomena consisted of case studies of small groups of people or sometimes only a single participant (see e.g., Milon, 1974). The structures being investigated occasionally would not appear in their test subjects' productions. Despite these limitations, the descriptions from these early studies provided rich insights into the theory of interlanguage. Still, others have called on the SLA research community to make better use of technology in order to further elucidate their findings:

> "Time has now come […] to test some of the current hypotheses on larger and better constructed datasets, as has happened in L1 acquisition. Not only do we need large datasets in order to be able to generalize our findings, but some of the structures which are crucial for informing current debates are rarely found in learner data. They therefore either must be elicited specifically, or large datasets are needed in order to maximize their chance of being present" (Myles, 2005, pg. 376).

As part of my thesis, I investigate whether the acquisition order of English negation emerges in ELL writing using a corpus of over 1 million word tokens. Others have also taken the challenge issued by Myles. In the sections that follow, I will review studies that have searched for an acquisition order using the methods outlined above for a variety of English structures including English negation, question forms, and the verb system. Some of these have been subsequently corroborated by more empirical, corpus-based research where others have not.

2.2.1 Acquisition of the English Verb System

Housen (2002) made use of an annotated corpus of learner English to corroborate previous research findings with regard to acquiring the English verb system. Additionally, he was concerned with what constitutes acquisition in the broader context, pointing to studies that considered mastery of the *form* of a structure as acquisition verses others that considered both its *form* and *function*. As such, he investigated the following questions: “how [do] learners acquire these basic forms; what stages of development can be seen in their acquisition; how [do] L2 learners map these forms onto their appropriate temporal, aspectual and grammatical meanings, and what stages can be observed in the development of these form-meaning relations” (pg. 77). To find answers, he analyzed the interlanguage of 46 ELLs from Dutch and French L1 backgrounds (23 apiece) from the Corpus of Young Learner Interlanguage. Interviews were recorded and transcribed at intervals over three years, amounting to a corpus of about 230,000 words. Table 2 presents what previous research via case studies had found up to Housen’s corpus study.

Stage	**Verb form**
1	Present Participle (Ving)
	Irregular Present of copula Be (*am*, *is*, *are*)
2	Progressive Aux/Be + Ving
	Irregular Preterit of copula Be (*was*, *were*)

3	Irregular Preterit (Ven) of lexical verbs
4	Regular Preterit (Ved) of lexical verbs
5	Regular Present (Vs) of lexical verbs
6	Irregular Present (*does*, *has*)
	Present Perfect Aux Have + Ven/ed

Table 2: Hierarchy of development of verb morphemes in English-L2 acquisition

Housen's data showed an overall pattern of underuse of these structures at the lower proficiency levels, followed by overuse, and then correct use as proficiency increased. This U-shaped behavior is commonly found in SLA research and is believed to show the interrelationship of form and function of a particular structure as it is being acquired (Ortega, 2014). This prompted Housen to organize the acquisition order of English verbal morphemes into three broad developmental stages which considered both their form and function. He also noted that the order of acquiring the form of verbal morphology generally agrees with the one shown in Table 2, however, he argued for characterizing the totality of English verbal acquisition in the following way.

Stage 1 is characterized by the use of invariant forms of verbs. This includes unmarked base forms regardless of any inflectional requirements induced by syntax such as *he see me yesterday* and highly frequent past and present participles like *got* or *running*.

Stage 2 is where some morphological adaptation takes place emerging progressively in the order: Vø > Ving; was > Ven > Ved; going + Vinf > have + V; Vs; will + V. Use of the base form in stage 1 is generally followed by the acquisition of the present participle, followed by the past participle, past tense and infinitive form. Some examples of stage 2 development are provided below in (1) and (2).

(1) [the interviewer asks the informant whether her friend speaks any French]

LIf6: no uh ... she speaking uh Nederlands.

(2) [the informant is asked to describe a picture of a man falling from a ladder]

INV: and what is he doing here?

LIf6: uh she fall.

Stage 3 is where verb forms become increasingly target-like with agreement, tense, and aspectual morphology attached.

Housen concludes with a call for more investigations into interlanguage development to corroborate and/or further elucidate sequences of acquisition using longitudinal data from individual learners. The present thesis makes use of a quasi-longitudinal corpus of learner data, which, according to Granger (2004), is the most common approach to accomplishing these aims. While there are some key methodological differences between Housen's study and the present one, the goal is the same. Housen (2002) articulates this common goal in his conclusion, "The combination of a substantive annotated computer corpus […] made it possible to empirically validate previous research findings obtained from smaller transcripts, as well as to test explanatory hypotheses about pace-setting factors in second language acquisition" (pg. 108). The composite picture Housen discovered will be considered in analyzing the acquisition order of English negation from my corpus. This will inform the engineering of the features related to English negation for use in an AES system.

2.2.2 Acquisition of English Interrogative Form

Pienemann et al. (1988) were the first to uncover an acquisition order for interrogative constructions in English (see Table 3). They found that novice ELLs first indicate they are asking a question by the use of rising intonation (Stages 1 and 2). They then advance to fronting a *wh*-word or an auxiliary without subject-auxiliary inversion (Stage 3). Next, inversion beings to take place in *wh*-questions with the copula and in *yes*/*no* questions with all auxiliaries except *do* (Stage 4). Finally, in the last two stages, inversion expands to the full range of possible contexts with target-like constructions and special cases emerging in the final stage (Stages 5 and 6).

Stage	Description	Illustration
1	Words and fragments with rising intonation	*A ball or a shoe?*
2	Canonical word order with rising intonation	*He have two house in the front?* *The boy threw a shoes?*
3	Fronting of a question element (*wh*-word, *do*, something else)	*Where the little children are?* *What the boy with the black short throw?* *Do the boy is beside the bus?* *Is the boy is beside the bus?*
4	Inversion in two restricted contexts: (1) In *wh*-questions with copula (2) In *yes*/*no* questions with auxiliaries other than *do*	(1) *Where is the sun?* (2) *The ball is it in the grass or in the sky?* *Is there a dog on the house?*
5	Inversion expands to the full range of target-like contexts	*How many astronauts do you have?* *What is the boy throwing?*
6	Negative questions Question tags Questions in embedded clauses	*Doesn't your wife speak English?* *You live here, don't you?* *Can you tell me where the station is?*

Table 3: The emergence of questions in L2 English (Ortega, 2014)

Spada & Lightbown (1999) added to this area of research by noting that some L1-influenced sub-stages may also exist. They analyzed acceptability judgements for English interrogatives by 144 French learners of English, discovering that because some kinds of *wh*-

questions in French do not require inversion where equivalent English questions do, these learners were more likely to accept certain ungrammatical English constructions. Thus, they were exhibiting a unique interlanguage stage conditioned by their L1.

To my knowledge, this acquisition sequence has yet to be corroborated by a longitudinal or quasi-longitudinal study as recommended by Myles (2005).

2.2.3 Acquisition of English Negation

Language acquisition researchers and developmental psychologists have long been interested in the development patterns of the L1 in children (see e.g., Klima & Bellugi, 1966). Beginning in the 1970s, SLA researchers began to apply what had been learned about the acquisition order of negation in L1 English to L2 English. Milon (1974) studied the English negation usage of "Ken", a seven year old Japanese boy whose family relocated to Hawaii, over a period of six months finding that his path of acquisition was congruent with that of native English speakers. Schumann (1979) sought to uncover if there were any unique characteristics of the acquisition of English negation for L1 speakers of Spanish. In doing so, he reviewed studies of the acquisition of English negation that included participants from seven different L1 backgrounds (Japanese, French, German, Norwegian, Taiwanese, Greek, and Italian). By comparing what researchers had found in the various L1s, Schumann outlined what amounted to a common sequence made up of four stages of development.

There was some disagreement early on about the structures included in the acquisition order and their sequence. Eskildsen (2012) explains that with so few participants, it may have been possible that researchers were overgeneralizing a finding that was erroneous based on the natural oscillation of acquisition. As discussed in 2.1, emergentists would argue that some oscillation is expected as non-target-like negation constructions are replaced overtime with more target-like ones. What is outlined in Table 4, however, has since been generally accepted and confirmed in SLA textbooks (Eskildsen, 2012).

Stage	**Verb form**
1	Internal negation – 'no', 'not' placed at beginning of clause: *No have money*
2	External negation – Negation placed between S and V: *She don't like me*
3	Auxiliary negation – Negation placed after auxiliary: *isn't, can't*
4	Analyzed do-negation – Target-like usage: *He doesn't laugh like us*

Table 4: Hierarchy of development of negation in English-L2 acquisition

Ahmad (2002) had 79 ELLs from different L1 backgrounds and proficiency levels judge the grammaticality of negation constructions represented in the four stages. To do so, she compiled a list of examples taken from prior research related to the L2 acquisition of English negation and determined where these manifestations would fall relative to one another within the acquisition hierarchy. Out of necessity, Ahmad separated each stage out so that the instances were only described by a single characteristic of the stage. For example, instead of stage 1 describing both pre-verbal 'no' and 'not', examples that included pre-verbal 'no' (*No have money*) and pre-verbal 'not' (*I not understand*) were placed into separate subcategories within stage 1. This provides a more precise view of the examples deemed acceptable by researchers to fit into this acquisition hierarchy. Having these precise definitions also makes comparisons of rank orders between studies more transparent (Dulay et al., 1982).

The results of the acceptability judgements revealed evidence of overlap between the stages as the learner progresses toward more target-like patterns and showed some differences in acquisition order based on the L1 of the learner, consistent with the previous studies reviewed in this chapter. Based on these results, there may exist a need to adjust the stages similar to what Housen did for the acquisition order of the English verb system. By looking at confidence intervals for the structures that make up the acquisition order of negation relative to the proficiency of the participants, one could definitively determine whether Ahmad's study and others were indeed seeing overlap between the stages or if the overlapping could be better explained as different manifestations of the same stage.

2.3 Written Response Rating

The scoring of written response items for high stakes testing traditionally requires the use of two human raters to achieve a reasonable level of reliability (H. Chen & He, 2013). Scores can either be derived holistically (a single grade for the essay as a whole) or analytically (separate scores for various aspects of writing). According to Weigle (2013) an analytic grading style is often used when assessing language through writing as in the case of ELL assessments. Scores are determined using a rating scale which describes characteristics associated with various levels of writing quality. Table 5 shows a template for a holistic scoring rubric adapted from Mertler (2001) as an example of what human raters consult to determine their scores. The rubric is meant to offer guidelines to help raters distinguish between those who possess the ability being assessed and those who do not. The purpose of the test will dictate the characteristics being analyzed and the ability test-takers must demonstrate at the successive rating bands. If the human raters do not agree and adjacent scores are given, the mean of the two scores is usually taken. This means that there are 11 possible scoring categories on a 0-5 rating scale like the one shown below. If the scores of the two raters are too far apart as determined by the test administrators, a third rater is brought in to arbitrate.

Score	Description
5	Demonstrates complete understanding of the problem. All requirements of task are included in response.
4	Demonstrates considerable understanding of the problem. All requirements of task are included.
3	Demonstrates partial understanding of the problem. Most requirements of task are included.
2	Demonstrates little understanding of the problem. Many requirements of task are missing.
1	Demonstrates no understanding of the problem.
0	No response/task not attempted.

Table 5: Template for Holistic Rubrics

One of the most advanced forms of human rating employed in educational industry today involves the use of Rasch modeling. The multi-faceted Rasch measurement (sometimes called the 'fair average') accounts for and removes "construct-irrelevant variance" such as individual rater severity and prompt difficulty (Coniam, 2009). This produces scores that are on a continuum rather than categorical in nature. For example, if one rater scored an essay as a 3 and another rater scored the essay as a 4, the score might come out to 3.2 instead of 3.5 if the rater who gave the 4 is shown by the model to be statistically more lenient than the average rater.

2.4 Automated Essay Scoring (AES)

Despite advancements in human scoring through Rasch modelling and other techniques, the time and monetary resources required to have humans score extended response items has caused AES systems to emerge as a desired alternative. Project Essay Grader, or PEG, was the first AES system developed in 1973 by Ellis Page (H. Chen & He, 2013). This system used a combination of features classified as simple, deceptively simple, and sophisticated to arrive at similar levels of agreement with human raters as agreement between only human raters, which at the time averaged to about .70 (Shermis, Burstein, & Bursky, 2013).

Simple features are those that directly relate to an aspect of writing quality that human raters look at when scoring essays. These aspects might even be explicitly stated in the scoring rubric. An example could be a count of adjectives as these can indicate to human raters a higher quality of writing. Deceptively simple features are those that are correlated with writing quality but are usually not looked at by human raters. An example might be the word count of the essay, which has been shown to be used by human raters up to a certain point until the essay length is deemed sufficient, at which point other aspects of the writing become more important. Sophisticated features are those that indirectly relate to an aspect of writing quality or proficiency that human raters look at when scoring essays. Counts of transition words may indicate text cohesion to an AES system where a human rater would almost never consider counting these words to determine cohesiveness. AES systems today still use a variety of features from these same categories, although an emphasis is being placed on the need for more sophisticated measures and less simple and deceptively simple ones (see e.g., Chodorow & Burstein, 2004; Weigle, 2013; Vajjala, 2017).

Most AES systems used in educational industry today are built using natural language processing modules combined with machine learning algorithms. As Figure 1 shows, these systems are trained on a set of human-rated, or “labeled”, essays by extracting features from the writing that are thought to be predictive of the skill or skills being assessed. Once the system reaches a high level of predictive accuracy on a training set, a model is created, which can then be used to score “unlabeled” essays.

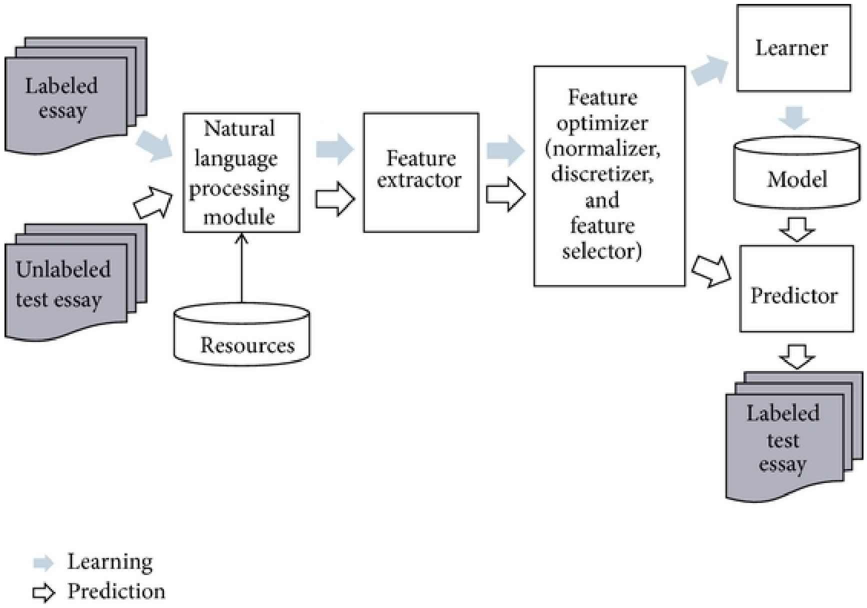

Figure 1: Automated essay scoring system architecture based on machine learning (Yi, Lee, & Rim, 2015, Figure 2)

2.4.1 Training Data

It is imperative that conditions match between the essays that the AES system will be trained on and those that it will eventually score. For example, a scorer that was trained on 30-minute essays is not likely to give an accurate predicted score for a 10-minute essay. Likewise, an AES system trained on essays scored by humans on a 0-5 scale will not predict scores outside of that range unless hand-coded to do so.

Similarly, the purpose for which the training essays were written must match the purpose of the essays to be scored by the AES system. Weigle (2013) describes three possible purposes for writing tests as outlined below:

(1) Assessing writing (AW)—Does the student have skills in text production and revision, knowledge of genre conventions, and an understanding of how to address readers' expectations in writing?

(2) Assessing content through writing (ACW)—Does the student understand (and display knowledge in writing about) specific content?

(3) Assessing language through writing (ALW)—Has the student mastered the second language skills necessary for achieving their rhetorical goals in English?

ALW tests often assess writing ability and content in addition to language either by design (e.g., aspects related to writing ability and content are explicitly listed in the scoring guide) or inadvertently through rater error. This can complicate matters if an AES system built strictly for ALW essays was trained on essays that assessed the writer's rhetorical style and written content as well.

The amount of training data, or scored essays, needed to reach a high level of agreement with human raters also depends on the purpose of the test. An AES system built for ACW tests may initially require less training data than AW and ALW tests because the prompt naturally constrains the content of the responses, but these systems must be trained separately on each prompt. Dikli (2006) surveyed a number of the most popular AES systems built for ACW tests and found that most require around 300-500 training essays per prompt. AES systems for AW and ALW tests will likely require more training data because there are many ways to demonstrate what is being assessed—rhetorical strategies and language proficiency, respectively—and thus a greater range of possibilities for the machine learning model to account for. Prompt-agnostic systems may have the advantage in this regard even though they initially require more training data to build.

Generally speaking, the more essays a system is trained on, the more reliable it will be, so long as their conditions match those of the essays that the AES system will score (Coniam, 2009). In the past, the need to obtain large sets of training data was a major criticism of AES (Dikli, 2006). For AES in ELL contexts, Lonsdale & Strong-Krause (2003) built a system by training on roughly 300 essays and came within one point of human scores on a 5-point scale 66% of the time. Millett (2006) extracted additional features from the same 300 essays improving agreement with human scorers to 90%, but still recommended using a larger corpus in future research. Yannakoudakis et al. (2011) recognized the need for a large, publicly shared dataset against which multiple studies related to AES for learner essays could be conducted and compared. Their dataset called the CLC FCE Dataset[1] contains 1,244 essays produced by ELLs from all over the world. This corpus can be downloaded online and represents significant progress in the field.

2.4.2 Feature Selection

While the training data must be sufficiently large for the prediction model to be accurate, the features also need to be carefully selected based on the purpose and scope of the test. Though AES systems like those mentioned above have been around for decades, there still is not a standard set of features found to be most predictive of scores for a variety of written contexts. The generalizability of many features also has not been well documented across datasets because most AES systems are developed only for proprietary purposes (Dikli & Bleyle, 2014; Vajjala, 2017). A common starting place for deciding which features to extract from the essays is the scoring rubric human raters use to determine scores (Briscoe et al., 2010). Encoding the

[1] https://ilexir.co.uk/datasets/index.html

characteristics found in the scoring rubric can be challenging but ultimately offer more meaningful feedback than some of the more superficial features used in many AES systems. This also improves the validity of the scores. Attali (2013) reports that validity is: “the degree to which evidence and theory support the interpretation of the test scores entailed by proposed uses” (pg. 181). Ultimately, accuracy with the given dataset will most often dictate which features are included in an AES system, regardless of whether they require deep linguistic processing (Vajjala, 2017), but this should be done with caution as it has an effect on the validity of the scores the system produces.

The most commonly used features in AES systems tend to be the simple and deceptively simple ones such as various counts of characters, words, and sentences because of their high correlation to scores and their relative ease to encode. These kinds of counts are said to measure the organization and development of an essay (Attali, 2007; Weigle, 2013). Additionally, deceptively simple measurements like average word length can indirectly show the complexity of the essay since longer words tend to be less frequent than shorter ones and often require affixation (Chodorow & Burstein, 2004; Yannakoudakis & Briscoe, 2012). Length features tend to correlate especially high with scores on learner essays because a person with greater language proficiency is likely to be able to produce longer words, sentences, and a longer essay overall in the time allotted than those who are not at the same proficiency level (Chodorow & Burstein, 2004; Zesch, Wojatzki, & Scholten-Akoun, 2015; Vajjala, 2017).

The use of length features, particularly essay length, has drawn the majority of the criticism AES has received because most human reader scoring guides make no direct mention of them. In their report on the use of *e-rater* on non-native learner essays, Burstein & Chodorow (1999) discussed the decision to remove any variable that is not explicitly part of most human reader scoring rubrics, including essay length. They conceded, however, that inherent in the requirement to properly form and defend a thesis statement, greater essay length is often a byproduct. In some more recent cases, essay length *is* explicitly listed on scoring guides for non-native essays, meaning the measurement of essay length directly relates to what human scorers look at to determine writing proficiency. Of note, the ACTFL proficiency scale measures various lengths of utterances and written productions to denote language proficiency (ACTFL, 2012).

Attali (2007) also acknowledged the natural side effect of a greater essay length when providing the detail called for in a scoring rubric used for native essays. In doing so, he recognized that multiple-choice tests of writing similarly provide an indirect view of the learner's writing ability, but he maintained that it still negatively impacts the validity of AES because it is construct-irrelevant. A review of several studies that performed subtest correlations on writing tests with both selected response and extended production items revealed that correlations between human-scored essays and scores on different subtests were, in fact, slightly higher than correlations between automated essay scores and other scores.

Despite the criticisms, essay length remains an important feature found in most AES systems. This has led Weigle (2013) and others (see e.g., Vajjala, 2017) to justify its use while recommending the continued search for and implementation of more sophisticated features. Chodorow & Burstein (2004) found that removing essay length from *e-rater* decreased the variance of scores with human-raters significantly, causing them to state, "the greatest improvements in machine scoring will come from the development of new features for as-yet unmeasured aspects of composition" (pg. 31).

Beyond simply taking raw counts of characters, words, and sentences in an essay, many linguistically interesting phenomena can be measured by employing these counts creatively.

Approaches involve counting the amount of sentences in a preset amount of words to measure sentence density (Millett, 2006), counting specific function and content words and certain parts of speech to measure cohesion (Lei, Man, & Ting, 2014; Crossley, Kyle, & McNamara, 2016), and counting clauses and T-units to look at the complexity of writing (Yang et al., 2015; Vajjala, 2017). The type-token ratio, which is derived by dividing the number of unique words by the total number of words in an essay is also a measurement found in most AES systems. Another commonly used feature extracts grammatical errors from the essay, though, surprisingly, these have not been found to be very predictive of scores (Millett, 2006; H. Chen & He, 2013). Zesch et al. (2015) applied readability features in their study including the Flesch, Coleman-Liau, ARI, Kincaid, FOG, Lix, and SMOG measures. McNamara, Crossley, & McCarthy (2010) used the online tool Coh-Metrix, which measures text cohesiveness via features related to discourse connectives, coreference, sentence overlap, among others, to determine if cohesion was predictive of writing proficiency. Additionally, Yannakoudakis & Briscoe (2012) successfully modelled discourse coherence in ELL essays. The features related to cohesion and coherence in these studies improved the predictive accuracy of their respective baseline AES scorers.

2.4.3 AES Criticisms and Concerns

As touched on previously, the implementation of these systems in educational industry has been met with controversy since being implemented in more high-stakes testing as in the case of the Pearson subsidiary, Edexcel, which in 2009 announced its decision to grade the PTE Academic entirely by machine (Briscoe et al., 2010). The critiques of these systems generally revolve around the idea that machines do not truly understand essays as humans do and may miss "intrinsic variables of interest such as diction, fluency, and grammar" (Attali & Burstein, 2006, pg. 3). The goal behind the inclusion of any feature in an AES system, however, is always related to improving the validity and reliability of the system and attaining high levels of agreement with human raters. While the connection to actual writing quality may be indirect, Briscoe et al. (2010) reassure that many features can be viewed as proxies for understanding the essay. They conclude, "In the end, this type of philosophical objection tends to dissipate as algorithms become more effective at any given task" (pg. 11). This sentiment *has* dissipated in the case of some state of the art systems like *e-rater*, which predicts scores that correlate with human ratings as high as correlations between two human raters, which is in the .90s for trained professionals today (Attali & Burstein, 2006). Another way to mitigate these concerns is by employing the AES system only in conjunction with another human rater, thereby gaining some of the financial and time-saving benefits of automated scoring while still retaining the human element. This is how the essay portions of the GRE, GMAT, and TOEFL are scored (H. Chen & He, 2013; Weigle, 2013; Williamson, 2013).

Another common concern about the use of AES systems is their susceptibility to being gamed. If a system doesn't include features based on more sophisticated natural language processing approaches and instead relies mostly on counts of words and sentences, these systems can be exploited by *keyboard banging*, *bad-faith*, or *unexpected topic* essays (Burstein, Tetreault, & Madnani, 2013). Keyboard banging essays are those that include a series of randomly typed characters possibly separated by spaces and punctuation marks to trick a system into considering them to be words and sentences. This would only work if the AES system relies *solely* on generic counts of words and sentences, which is rare even in low-stakes testing. A test-taker could, however, produce an actual coherent essay that includes a random string of characters to trick the scorer into predicting a higher score based on the overall length of the essay.

Bad-faith essays include coherent words and sentences but of inauthentic, memorized content such as song lyrics or poetry. Features related to detecting faithfulness to the prompt have been experimented with and successfully implemented in AES systems like *e-rater* (Burstein et al., 2013). This can be costly and cumbersome as the system would need to be retrained on each individual prompt to perform a content vector analysis. This takes the content of essays that demonstrated faithfulness to the prompt and ranks the words found therein in order to check how similar a new essay is compared to this corpus of faithful essays. Another, more obvious solution for detecting and appropriately scoring such essays would, again, be combining AES scoring with a human scorer. As Yannakoudakis et al. (2011) put it, "The practical utility of an [AES] system will depend strongly on its robustness to subversion by writers who understand something of its workings and attempt to exploit this to maximise their scores" (pg. 185). In doing so, the validity of the system is vastly increased.

Despite some of the well-documented advantages and advancements of AES, there is still pushback and likely will be until there is near 1-to-1 agreement with human scorers. One area where AES has an obvious advantage when compared to human scoring is in terms of reliability. Williamson (2009) tackled the argument in favor of using automated scoring despite the inevitable imperfections from this perspective citing common human inconsistencies in scoring due to halo effects, fatigue, and the overlooking of details. He postulates that these human-errors may be more of a contributing factor to the discrepancy between human and automated scoring and the reason why AES systems should be expected to produce scores that occasionally differ from human scores.

Regardless of the seeming inevitability of difference between human and automated scoring, numerous studies have called for more sophisticated features that connect closely to the scoring rubrics human raters use. Chodorow & Burstein (2004) state that “not only will this lead to better scoring performance, but it will also let us capture more of the richness and diversity of human language” (pg. 31). To that end, many of the above studies introduced new features as a means to investigate their usefulness in an AES system and their connection to writing quality that human raters look for. In this thesis, I attempt this with the use of a set of features based on the acquisition order of English negation.

Chapter 3 - Methodology

As stated previously, the foundational studies that looked into the phenomena of L2 acquisition orders were often built on observations of very few participants or on faulty methods. This sometimes resulted in the erroneous rank ordering of structures that were either too closely related or varied widely based on individual differences or general L1 influences. More recent research has sought to corroborate these foundational descriptions with larger sample sizes, longitudinal frameworks, and more empirically based evaluation methods.

This thesis features a large written corpus of quasi-longitudinal ELL data to investigate the claims made about the acquisition order of English negation. In this chapter I will discuss the corpus I chose for analysis, the method I used to extract instances of negation from the essays, and how I have chosen to evaluate the order and distinctiveness of each stage. Following this, I will discuss the AES system I built and its accuracy in predicting scores for a test set of essays without the use of features related to the acquisition order of English negation. I will then introduce the ways in which I operationalized the learners' negation usage for use in the AES system to judge whether these features positively contribute to the prediction of scores for ELL essays.

3.1 The Corpus

The corpus I used is comprised of 3,633 ELL essays written by students enrolled in Brigham Young University's English Language Center. There are a total of 1,144,281 word tokens in the corpus. This averages to about 315 tokens per essay with a range of a single token to 1,091 tokens. The average age of the students is just under 25 years old, ranging from 17 to

63. Female student essays account for 57% of the corpus. A total of 41 L1s are represented in the corpus with the top 10 provided in Table 6 and the full list supplied as 0.

L1	Essays
Spanish	1594
Korean	620
Portuguese	447
Chinese	314
Japanese	202
Russian	92
French	59
Ukrainian	38
Mongolian	34
Thai	31

Table 6: Top 10 L1s represented in the corpus

The mean length of study for students in this program is approximately three semesters, which means that there are multiple essays written by some of the same students at different intervals in the corpus. In fact, when removing duplicate student identification numbers, 1,786 unique ID numbers remain connected to the 3,633 essays. This corpus cannot be described as longitudinal, however, as no efforts were made to map the progress of each student individually over the course of their study. Each essay was treated as an independent writing sample since even the essays from the same individual show different growth intervals and were written for different prompts. Thus, this corpus fits better the description of a quasi-longitudinal dataset described by Granger (2004) as, "[a corpus] gathered at a single point in time but from learners of different proficiency levels" (pg. 131).

The essays were written during 17 semesters between Spring 2009 and Spring 2015. After being admitted, newly enrolled students in this program take a placement test (PT) with sections on grammar, reading, listening, speaking, and writing in order to be placed into one of the seven levels offered based on their proficiency. Continuing students take a similar test called a Language Achievement Test (LAT) which serves to adjust the student's overall proficiency level as needed at the end of each semester. A majority of the students fall in the Intermediate Mid to Intermediate High English proficiency level according to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines (ACTFL, 2012).

The written portion of the PT includes two prompts—one for a 10 minute essay and another for a 30 minute essay. Despite there being two prompts, a single score is allocated based on the student's writing performance, collectively. For consistency, I only used the 30 minute essays from the PT. The LAT written section consists of a single 30 minute essay prompt.

The prompts for the essays are different for each semester in the corpus. Because the goal of these assessments is to judge English proficiency, there isn't a heavy emphasis placed on the students' ability to adhere to the prompt. As such, I could only find one example of a student receiving a zero ostensibly for not adhering to the prompt. Examples of the kinds of prompts used for the essays are provided below:

a) You are the freshman counselor at a university. You are writing an essay to all incoming freshman that answers the following question: What are the 3 most important subjects that students should study in college to prepare them for the future? Why?

b) Identify one improvement that would make your hometown more appealing to people your age and explain why people your age would benefit from this change. Use specific reasons and examples to support your opinion and describe the potential immediate and long-term consequences of this improvement.

The score for the written portions of the tests is the average of two human ratings on a 0-7 scale. Prior to the fall 2010 semester, the average of these two ratings was taken for the score, but afterwards the Rasch-derived ‘fair average’ described previously was used. The descriptive statistics for the scores are provided in Table 7 and the amount of essays per scoring range are found in Table 8 below.

Mean	3.58
Standard Error	0.02
Median	3.61
Mode	3
Standard Deviation	1.29
Range	8
Minimum	0
Maximum	8 *
Count	3633

* Five essays scored above a 7 due to the Rasch model

Table 7: Descriptive statistics of scores

Range	**Essays**
0-1	67
1-2	259
2-3	724
3-4	1130
4-5	941
5-6	390
6+	122
Total	3633

Table 8: Essays per scoring range

To construct the corpus, I began with 4,025 essays, with 2,829 coming from the LAT and the remaining 1,196 from the 30 minute prompt on the PT. After removing any instances where there wasn't a score associated with an essay or the essay was left blank, I arrived at 2,483 LAT essays and 1,150 PT ones (totaling 3,633 essays). The reason for removing essays without a score relates to the need to build an AES system from this corpus. While what is written in these essays could be valuable for any instances of negation found in them, machine learning models can only be trained on labeled data, which in this case are essays that include scores. To ensure that I only analyzed the same essays that would eventually be used to build the AES system, these essays were removed from the corpus.

Blank essays with a score of zero were also removed from the corpus as they provide no information regarding negation usage and there is nothing else of linguistic significance to extract from them for a machine learning model to learn on. A blank essay with a corresponding zero could supply the model with an additional pattern for predicting a zero on new data, but this could just as simply be hand-coded prior to implementation if deemed necessary. Finally, there were a few instances where a student did not produce any writing and received a score greater than a zero. These were removed since they were likely instances where a student wrote something for the 10-minute prompt on the PT, which I did not use in this analysis, and nothing for the 30-minute prompt.

3.2 Developmental Pattern Extraction

I used a series of computational methods to extract examples related to the acquisition order of English negation from the essays in the corpus. To begin, I used the Python scripting language and the Natural Language Toolkit (NLTK) package[2] to tokenize the words in each essay and tag them with their part-of-speech (POS) labels. Tokenization is the process of

[2] http://www.nltk.org/

segmenting the words of a text, which would otherwise be viewed as a continuous string of characters by a machine. Tagging involves assigning POS tags to the tokens. NLTK uses the Penn Treebank tagset for this task (Marcus, 1993).

Once I tokenized and tagged the essays, I used regular expressions like the one depicted in Figure 2 to search for combinations of words and POS tags related to each of the stages of negation as introduced in Table 4. Regular expressions allow for intricate pattern matches within a body of text and are written using simple and special characters that can match more than the literal characters of the text. For example, in Figure 2, '\w+' will match one or more alphanumeric character.

```
\b((do)\W+\w+\W+(not|n\'t)\W+(RB)|(dont)\W{6}\w+\W+)\W{6}\w+\W+(VBG|VBD|VBZ|VBN|MD)\b
```

Figure 2: Regular expression for stage 2a

Table 9 lists each characteristic related to the stages of negation acquisition as described in the literature (see Ahmad, 2002), which I encoded and extracted from the corpus. I created sub-stages where there were multiple characteristics that acted as separate manifestations of a single stage (see sub-stages 1a and 1b). When a single regular expression could not extract every instance of a particular characteristic, I used multiple regular expressions and combined them into a single sub-stage (see sub-stages 1b and 4b). I extracted matches of these characteristics in the essays and output them to a spreadsheet next to the score of the essay from which they were extracted and the sentence that contained the instance for added context.

Stage	Characteristic	Instance
1a	Neg. particle *no* before V or negatum	*The people no understand the importance of careful of the live…*

1b	Neg. particle *not* before untensed V or negatum	*…a parent should not being called a good parent if she just give food.*
	Neg. particle *not* before tensed V or negatum	*I not forgot him.*
2a	Neg. element *don't* used but not marked for person, number or tense (pre-verbal *don't*)	*If we do not doing like that, we cannot be a great rommate.*
2b	Neg. element *don't* used before modals	*…we don't must to wait*
2c	Neg. element *doesn't* used but not marked for person, number or tense (pre-verbal *doesn't*)	*…they doesn't need pay nothing.*
3a	Neg. element *don't/doesn't/do not/does not* used after aux. verbs (*are*, *is*, etc.)	*People are don't like a person who say lie.*
3b	Neg. particle *no*, *not*, *n't* used after aux. verbs (*are*, *is*, etc.)	*…I haven't been there…*
3c	Neg. particle used after modals (*can*, etc.)	*…without education you can't have a good work.*
4a	Neg. element *don't* is marked for person, number or tense	*I don't regret my any decisions I have had.*
4b	Both aux. & V marked for 1SG, 2SG, 1PL, 2PL, 3PL or tense	*…we didn't had help*
	Both aux. & V marked for 3SG or tense	*It doesn't matter how far away we are from each other.*

Table 9: Stages of negation extracted from corpus

3.3 Confidence Intervals

As Dulay et al. (1982) emphasized, empirical methods need to be employed when describing an acquisition order so that the structures said to be acquired in a distinct order can be shown to be truly independent of one another. I used 95% confidence intervals to distinguish each characteristic listed in Table 9 relative to the score of the essay from which the instance was extracted. By doing so, I was able to analyze with a high degree of certainty the range of proficiency levels where the various structures surfaced.

3.4 The AES System

I built a fully functional AES system (AES1) to establish a baseline of predictive accuracy prior to introducing any features based on the acquisition order of English negation. This was done to determine whether encoding an acquisition order had any effect on the system's accuracy. I used various Python packages along with a machine learning algorithm called the Gradient Boosting Regressor from Scikit-learn[3] for this task.

The features I originally planned to extract to train up the scorer were those used by Millett (2006), who achieved 90% agreement with human scorers. These features included part of speech patterns, vocabulary density, along with word, sentence, and essay lengths. Millet based his research on previous work done by Lonsdale & Strong-Krause (2003) using a corpus from the same language center from which the corpus for this thesis was collected. Several of the features that Millet extracted, however, were based on specialized indices extracted by the text processor WordMap, which has since been acquired and its documentation removed from the public domain. Therefore, I experimented with nearly 200 features related to various aspects of ELL writing, finding the 40 listed in Table 10 to be the most salient to the corpus. They constitute various lexical, stylistic, developmental, and grammatical components of writing. The 31 features specifically related to the acquisition order of negation are listed beneath the grammar category and will be discussed subsequent to the 40 features that make up AES1.

Feature group	Features		
Lexical sophistication	WORD RANK TOTAL	AVERAGE WORD LENGTH	NUM_TYPES
	WORD RANK AVERAGE	TYPE TOKEN RATIO	NUM_SYLLABLES
	DIFFICULT WORDS		NUM_CHARACTERS

[3] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

Style, organization, development	NUM_SENTS AVERAGE SENT LENGTH SENT DENSITY TRANSITION WORDS NUM_TOKENS	Cohesion features NUM_DETERMINERS NUM_CONJUNCTIONS NOUN TYPE TOKENS RATIO PRONOUN DENSITY PRN-NOUN RATIO NLEMMAS NBIGRAM LEMMAS NTRIGRAM LEMMAS NLEMMA TYPES NBIGRAM LEMMA TYPES NTRIGRAM LEMMA TYPES NCONTENT TOKENS NCONTENT TYPES CONTENT TYPE TOKEN RATIO NFUNCTION TOKENS NFUNCTION TYPES FUNCTION TYPE TOKEN RATIO	Readability features FLESCH READING EASE FLESCH-KINCAID GRADE COLEMAN LIAU INDEX AUTOMATED READABILITY INDEX DALE-CHALL READABILITY SCORE LINSEAR WRITE FORMULA GUNNING FOG
Grammar	GRAMMAR CHECK	ENGLISH USAGE	RELEVANT POS TRIGRAMS
Negation	NEG_USAGE, S1A, S1B, S1C, S2A, S2B, S2C, S3A, S3B, S3C, S4A, S4B, S4C, S1, S2, S3, S4, NEG_USAGE_NEW, S1A_NEW, S1B_NEW, S1C_NEW, S2A_NEW, S2B_NEW, S3B_NEW, S3C_NEW, S4A_NEW, S4B_NEW, S4C_NEW, S1_NEW, S2_NEW, S3_NEW		

Table 10: Features extracted from the corpus

3.4.1 Features: Lexical Sophistication

The lexical sophistication features pull out salient information at the word level. Some of the features in this category look at the sophistication of individual words (WORD RANK LOCAL, WORD RANK TOTAL, DIFFICULT WORDS) and others analyze word choices globally (AVERAGE WORD LENGTH, TYPE TOKEN RATIO, NUM_TYPES, etc.). The DIFFICULT WORDS feature extracts and generates a raw count of the difficult words found in the essay according to the freely available Python package *textstat*[4]. The features WORD RANK TOTAL and WORD RANK AVERAGE both analyze the ranking of the content words in the essay relative to the top 5,000 words found in the Corpus of Contemporary American English (COCA)[5]. For WORD RANK TOTAL, the sum of the

[4] https://github.com/shivam5992/textstat

[5] https://www.wordfrequency.info/

rankings is taken and for WORD RANK AVERAGE, the sum is divided by the total number of ranked words found in the essay.

3.4.2 Features: Style, Organization, and Development

The features that stand as proxies for important aspects of style, organization, and development are found at the sentence, paragraph, and total essay level. The features NUM_SENTS, AVERAGE SENT LENGTH, and SENT DENSITY look at the total number of sentences in the essay, their average length, and how many sentences there are per 100 words to calculate density. The feature TRANSITION WORDS looks for any transitions in the essay as these are indicative of the proper organization and development of a thesis. The NUM_TOKENS feature gives the word count of the essay again related to its overall development.

I chose the cohesion features for the system by using the Tool for the Automatic Analysis of Cohesion (TAACO) on a subset of 1800 essays. This open source program developed by Crossley et al. (2016) can extract up to 150 indices related to cohesiveness in texts. I extracted these TAACO measures from the subset of essays, built a machine learning model to predict essay scores based just on these metrics, and encoded the top 10% most predictive ones in the AES system.

The features in the system related to the readability of the essays were also generated by *textstat*. Because the Simple Measure of Gobbledygook (SMOG) grade requires at least 30 sentences to be statistically valid, it was not included in this system as this would exclude a large portion of the corpus.

3.4.3 Features: Grammar

The features related to grammar focus on the rules and conventions specific to the English language. The GRAMMAR CHECK feature uses the Python wrapper for LanguageTool[6], which checks documents for over 2,000 unique error types. Once LanguageTool finds the errors in the essay, GRAMMAR CHECK counts them and divides this number by the essay's word count. ENGLISH USAGE ensures that the essays are written in English by getting the percent of the essay's words that are found in a large English wordlist.

I created the RELEVANT POS TRIGRAMS feature extractor based on a similar feature described in Millett (2006). In his thesis, he found that the percentage of WordMap's 84 specialized POS trigram patterns that appeared in an essay showed a significant positive correlation to its score. Based on this idea, I extracted each POS trigram pattern found in the corpus, built a machine learning model to determine which were the most predictive of scores, and used this set of 103 relevant trigrams to determine what percent were found in each essay. While not as salient as it was for Millet's study, this feature proved predictive enough to make it in the final cut.

The set of 31 features based on the acquisition order of negation fit into the grammar category. I used the regular expressions described in 3.2 to extract instances of each stage of negation represented in the essay. As various interlanguage forms characterize a single stage, I first extracted and counted each of these separately (e.g., S1A, S1B, S2A), then I added each sub-stage together to get a count of instances for each full stage in the essay (S1, S2, S3, S4). Next I added up all the instances of negation from any stage in the essay and called this feature NEGATION USAGE, and divided the count of each sub-stage by the total instances of negation to

[6] https://github.com/myint/language-check

determine what proportion of their negation usage fell into each sub-stage. Finally, I repeated this process for a newly ordered negation acquisition sequence that is based on the analysis in the next chapter. These features follow the same naming convention as those for the canonical order but are followed by _NEW.

Chapter 4 - Results

4.1 Research Question 1: Negation Acquisition in ELL Writing

In this thesis, I investigated whether the acquisition order of English negation surfaces in ELL writing. By using the above methods to extract examples related to each distinct characteristic described in the acquisition sequence, I found over 7,300 instances in the corpus. This is evidence that the patterns related to the acquisition order of English negation surface in ELL writing. To determine whether they followed the canonical order, I examined the amount of times the instances emerged at the different scoring ranges.

Table 11 shows the number of instances of each stage found in the various scoring ranges with the greatest amount highlighted for each stage. Evidence of the canonical acquisition order can be seen as there are no stage 1 and 2 constructions in the 6+ scoring range and relatively few stage 3 and 4 instances at the lower levels compared to the higher levels.

Range	**Instances**			
	STAGE 1	STAGE 2	STAGE 3	STAGE 4
0-1	2	1	12	7
1-2	10	10	118	94
2-3	37	31	672	418
3-4	**50**	**36**	1534	**744**
4-5	42	21	**1568**	657
5-6	9	5	706	248
6+	0	0	261	87
Total	150	104	4871	2255

Table 11: Number of instances of a given stage per scoring range

Additionally, this table shows that the total number of stage 1 and 2 instances is vastly underrepresented compared to the amount of stage 3 and 4 instances. This is likely caused by two main factors: (1) the unequal amount of essays per scoring range, and (2) the relative frequency of some of the structures in the stages being incongruent. For example, the number of structures described in stage 3b (negative particle *no*, *not*, *n't* used after auxiliary verbs) is far greater than those stages that merely look at the use of *don't* (see e.g., stage 2a).

In connection to these two factors, several smaller factors may contribute to the difference in representativeness of the stages in the corpus. First, essays that contain stage 1 and 2 forms might be expected to be shorter since they appear to be produced by learners at a lower proficiency level. If a learner is able to use stage 3 and 4 negation constructions, they might likely be at a higher level of proficiency and able to write more in the allotted time, producing a greater volume of these instances overall.

Lastly, the nature of stage 1 and 2 constructions technically being errors, where many of the structures in stages 3 and 4 contain correct forms, could have skewed the data in favor of the correct forms. If an ELL is making stage 1 constructions, they might not have a handle on spelling either, so there could be more occurrences of stage 1 and 2 structures that the regular expressions did not find due to errant tagging or misspellings. It would follow that there are probably more of these instances.

In the following section, the two main factors likely causing the varying representation are addressed and mitigated using several mathematical methods. The goal of this level of analysis is to enhance the visualizability of the distribution of the stages in the corpus.

4.1.1 Visualizing the Data

If the amount of essays per scoring range is normalized to 1,000 instead of the actual numbers which were very low at the extremes (see Table 8), advancements along the acquisition sequence are more clearly illustrated. As Table 12 reveals, stage 1 and 2 structures were used most by those in the 2 to 3 scoring range, where none of these structures were used by learners in the 6+ range. Conversely, stage 3 and 4 structures appeared most in writing at the 6+ level, with comparatively few instances in the 0 to 1 range.

Range	**Normed Instances**			
	STAGE 1	STAGE 2	STAGE 3	STAGE 4
0-1	30	15	179	104
1-2	39	39	456	363
2-3	**51**	**43**	927	577
3-4	44	32	1358	658
4-5	45	22	1666	698
5-6	23	13	1810	636
6+	0	0	**2175**	**725**
Total	231	163	8571	3761

Table 12: Number of instances of a given stage per scoring range (1,000 essays per range)

Further, to counteract the issue of ordering an incongruent number of structures per stage, Figure 3 shows the amount of instances in each scoring range if the total number of instances was 1,000 for each stage. This normalizes the representativeness of the structures and shows more clearly how stage 1 and 2 negation acquisition structures surface more at the lower scoring ranges before being overtaken by stage 3 and 4 structures at the higher levels. This also shows the global overlap between stages as less target-like constructions are replaced with more target-like ones as L2 English writing proficiency increases. This mostly corroborates the findings of Eskildsen (2012) who found “no acquisitional stage-defining pattern-dominance” (p. 30), except at the 6+ range where no stage 1 or 2 constructions were found. Finally, this shows that the canonical acquisition order of negation does in fact surface in the corpus.

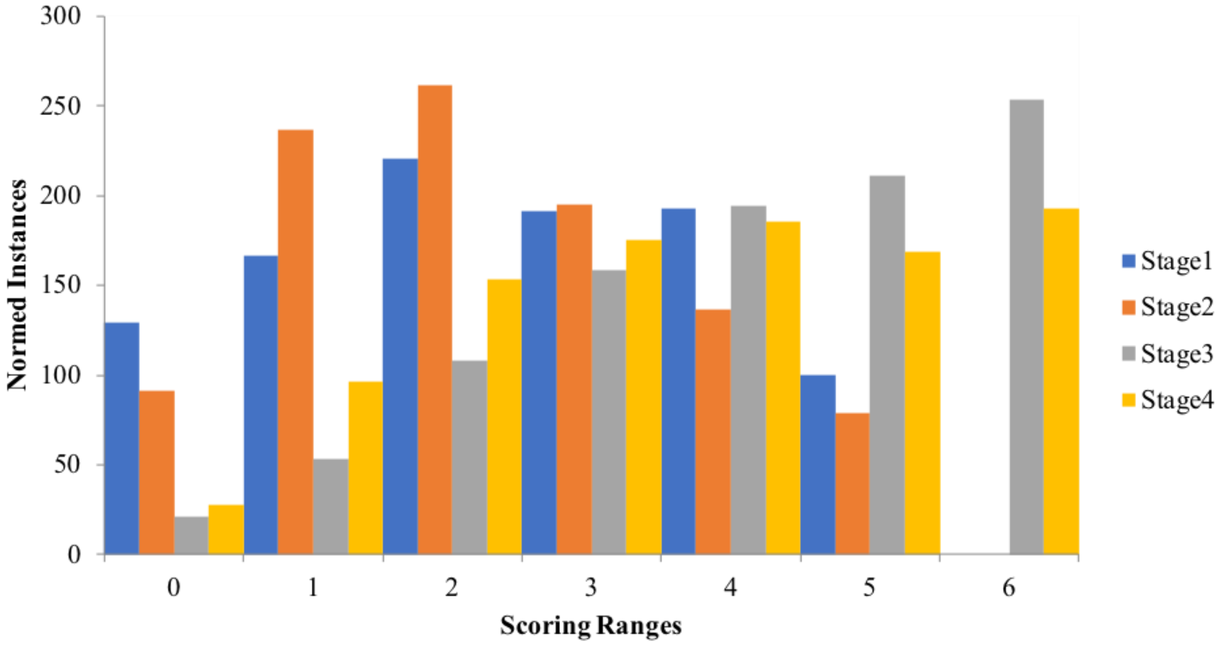

Figure 3: Histogram featuring normed instances of negation usage (1,000 total instances per stage)

4.1.2 Sub-stage Analysis

The four stages are broken down by sub-stage in Table 13, which also shows the average score of the essays from which the instances of these structures were extracted. By viewing the instances of sub-stage characteristics relative to mean scores, I can determine whether each characteristic is truly a manifestation of its stage as evidenced by it appearing at a similar scoring range as the stage's other characteristics. I can also see whether each characteristic falls in the canonical order when aligned with the written proficiency scores of the learners.

Sub-stage	Instances	Mean Score
1a	56	3.10
1b	94	3.58
2a	52	3.27
2b	43	3.17
2c	9	3.37
3a	16	3.32

3b	2642	4.02
3c	2213	3.94
4a	2077	3.71
4b	178	4.12

Table 13: Instances of negation at the sub-stage level

As can be seen, many of the average scores associated with the sub-stages follow a sequential order from low to high. Some, like 1b (negative particle *not* before V or negatum), however, seem to be noticeably misplaced in the sequence. Figure 4 further elucidates which of these sub-stages appear to be out of order in terms of the average score of the essay from which they were extracted.

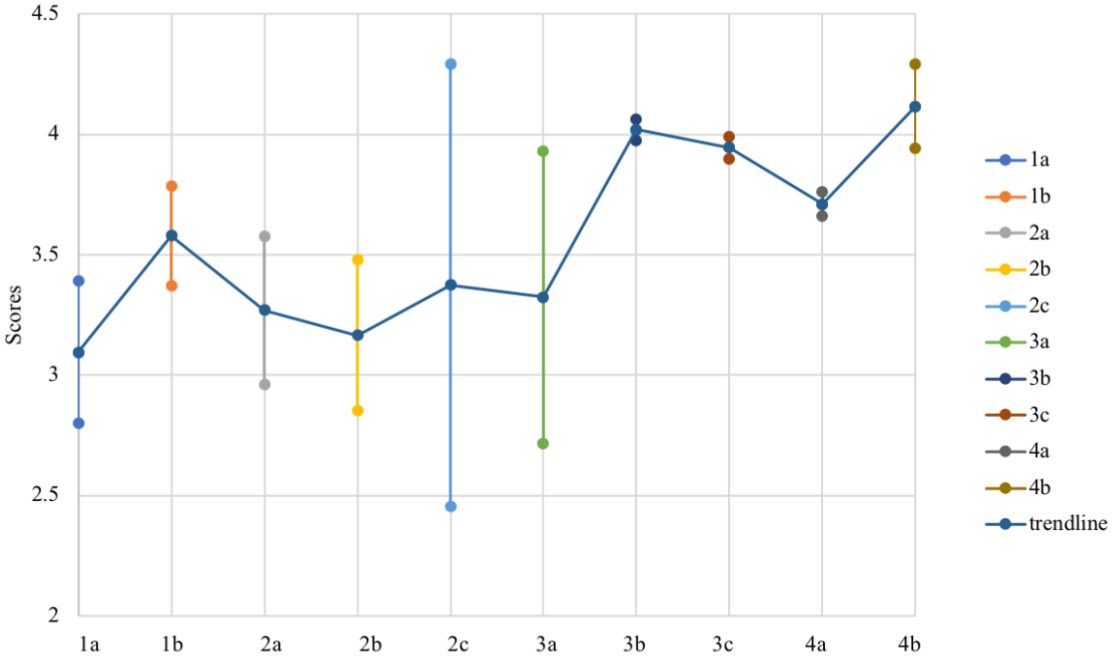

Figure 4: Confidence intervals for negation sub-stages

Several of the sub-stages such as 2c (negative element *doesn't* used but not marked for person, number or tense) and 3a (negative element *don't/doesn't/do not/does not* used after auxiliary verbs) contained too few instances to be reasonably confident about their order in this chart. Others, like 3b (negative particle *no*, *not*, *n't* used after auxiliary verbs), 3c (negative particle used after modals), and 4a (negative element *don't* marked for person, number or tense) can be declared distinct from any of the sub-stages outside of their intervals with 95% confidence according to this dataset. This is not to say that individual ELLs who receive lower written scores than the average of where these structures surface cannot produce stage 3b, 3c, or 4a structures—Table 11 clearly shows that these surface in all scoring ranges—but it definitively shows that those who successfully use these more target-like structures have a higher written proficiency score on average than those who do not.

Figure 5 proposes a reordered sequence of the sub-stages along the written proficiency scale. Note the changes to the order along the horizontal axis.

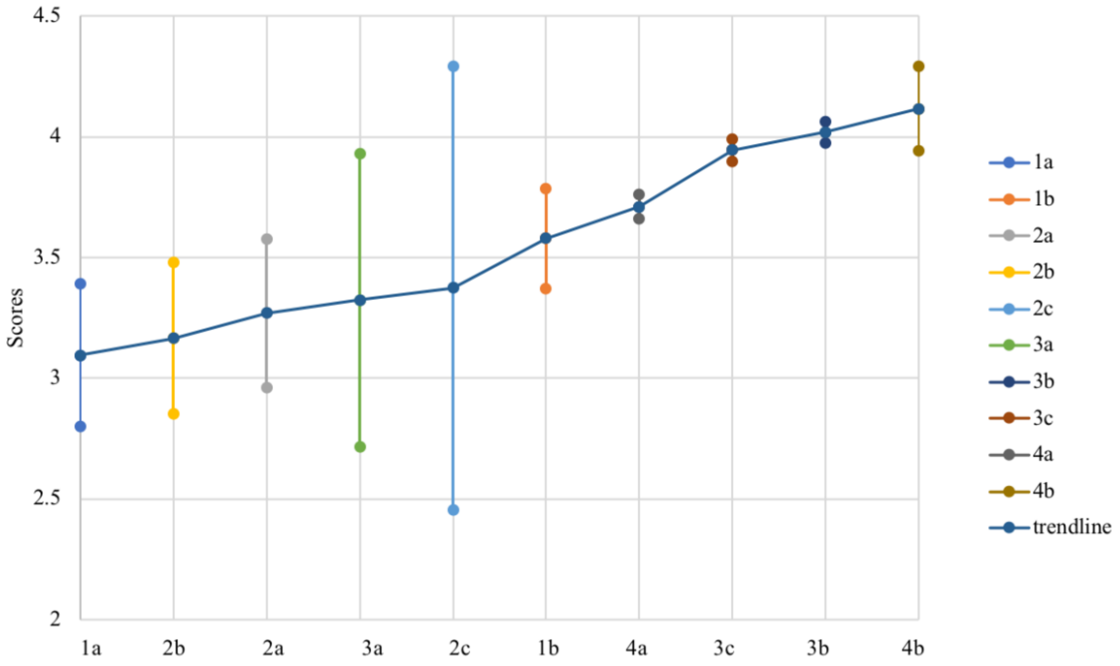

Figure 5: Reordered confidence intervals for negation sub-stages

Upon removing stages 2c and 3a from the analysis for paucity of data, three stages of acquisition can be reasonably extrapolated from the chart:

<u>Stage 1</u>
- 1a - Negative particle *no* before V or negatum
- 2b - Negative element *don't* used before modals
- 2a - Negative element *don't* used but not marked for person, number or tense

<u>Stage 2</u>
- 1b - Negative particle *not* before V or negatum
- 4a - Negative element *don't* marked for person, number or tense

<u>Stage 3</u>
- 3c - Negative particle used after modals
- 3b - Negative particle *no*, *not*, *n't* used after auxiliary verbs
- 4b - Both auxiliary and V marked for 1SG, 2SG, 3SG, 1PL, 2PL, 3PL or tense

Figure 6: A composite negation acquisition hierarchy in the written domain

Both the canonical acquisition sequence and this newly ordered one were encoded for comparison with the baseline system, AES1. This will be discussed in the following section.

4.2 Research Question 2: Using Features Related to Acquisition Order in an AES System

I used a regression model for the machine learning task as opposed to a classifier as many AES systems use because the labels (scores) in this corpus are continuous, floating-point numbers rather than categorical. When essays are scored by humans raters, traditionally the average of two human scores is given, and if the scores are more than 1-point apart, another human score is used to resolve the inconsistency. This naturally creates distinct scoring categories that a classifying model simply must *identify* as the best fit for unseen data. A regression model, on the other hand, is used to *predict* a label for unseen data that may have never emerged in the training data.

Because prompt difficulty has been shown to have an effect on writing quality (Yang et al., 2015), I first employed a 12-fold cross-validation method to train and test the scoring model of AES1. This technique involves shuffling the corpus randomly and partitioning it into folds. The system is then trained on 11 of the folds and tested on the remaining one, shuffled and retrained until it has been trained on each fold thus reducing the power of any one prompt. This also works to reduce the effect that a prompt that naturally elicits more negation might have on the system.

After training the model using the cross-validation split, the mean absolute error of AES1 on the test set of essays came to 0.5. This means that the scorer averages predictions that come within 0.5-points of the human score. Breaking this down further, the baseline machine gets within 1-point of the human score 90% of the time and 0.5-points 63.5% of the time.

With the addition of the 31 features based on the acquisition order of English negation in what will be referred to as AES2, the scorer's mean absolute error dropped slightly to 0.498, getting within 1-point of human scores 90.2% of the time and within 0.5-points 63.3% of the time. Thus, the scorer's predictive accuracy improved overall in several aspects and dropped slightly in its ability to predict scores within 0.5-points of human scores using this training and testing method.

Next, I employed an 80%/20% training/testing random split to train and test an additional scoring model using the same regression algorithm. Experimentation with various machine learning algorithms, parameters, and training and testing splits is common and necessary to achieve competitive results with supervised learning tasks. The random 80/20 split is a frequently used partition, which is why I experimented with it here. Table 14 shows greater improvement on all metrics with AES2 system scoring using the 80/20 split. MAE is down, and percent correct guesses on all ranges went up.

	AES1	AES2
MAE	0.4811	0.4803
Pct. exact	11.97	13.2
Within 0.5	66.57	66.85
Within 1.0	89.96	90.51
R^2	0.7857	0.7873

Table 14: Scoring results using 80/20 split

Figure 7 shows the distribution of AES2's predicted scores compared to the distribution of actual human scores. What is evident is that the scorer has more difficulty predicting scores at the extremes; particularly in the higher range. This is caused by the relative sparseness of essays at these scoring levels in the corpus.

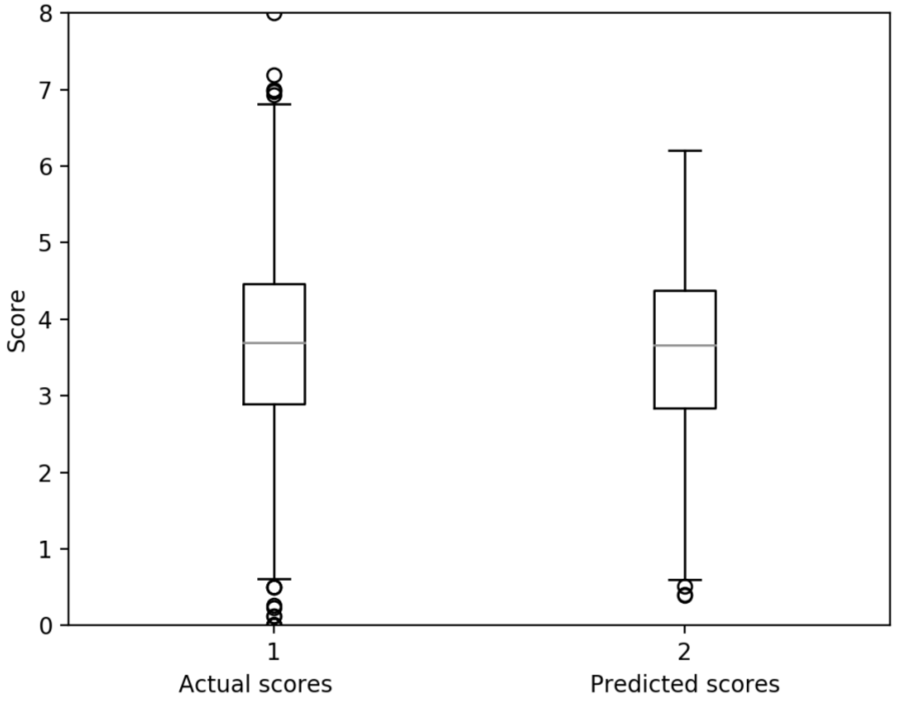

Figure 7: Boxplot featuring the AES scorer's predictive accuracy

The coefficient of determination (R^2) for AES1 was 78.57% and rose slightly to 78.73% with the addition of the negation features. Figure 8 is the graphical representation of this measurement and illustrates the variance. Again, the predicted scores tend to fall further away from the actual scores at the higher and lower scoring ranges. Specifically, the scorer tends to predict higher scores than the human rating at the lower levels and lower scores than the human rating at the high levels.

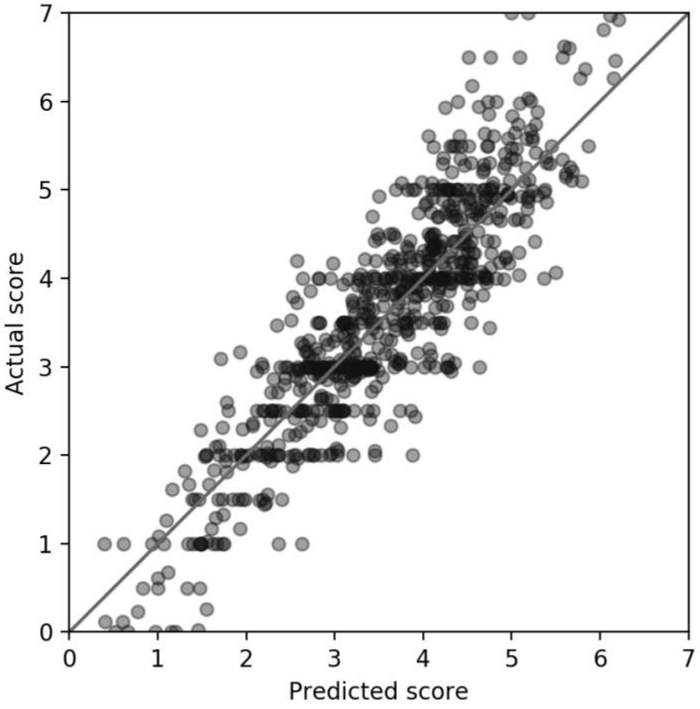

Figure 8: Scatterplot of predicted scores relative to human scores

Note that the evaluation methods employed here are not those commonly used in assessing the reliability of AES systems (Yannakoudakis & Cummins, 2015). Typically the reliability of an AES system is based on a variety of metrics that compare the machine-human scoring discrepancy to the human-human scoring correlation. Thus, if the machine-human scoring correlation is low, it does not necessarily mean that the AES system is unreliable unless it is significantly worse than the human-human correlation. Because human raters are still the gold standard that AES systems are trying to replicate, it is only important that the machine scores are consistent with the human-human correlation or perform better.

The database from which I collected the corpus for this thesis unfortunately only included the average score of the two human ratings, and thus the standard methodologies could not be employed. While not ideal, adjacent agreement in the 90% range *is* competitive with current industry standards. Also, the metrics employed in this section were used more as a means to compare two AES systems, AES1 and AES2, and not an AES system with the human standard. AES2 showing improvements in the measures used adequately answers my second research question about the effect of including features based on the acquisition order of negation in a fully functional AES system.

Chapter 5 - Discussion and Conclusion

In this thesis I encoded features based on L2 English acquisition patterns to determine if they could be useful in an existing AES system that achieves competitive results. This is an important question because researchers and assessment specialists have begun to challenge the validity of AES and question how the scores they produce translate to the target-language use domain. Previous studies that have undertaken to prove the reliability of AES by showing levels of agreement with expert raters are crucially important and more such studies should be conducted to maintain the argument that AES can be a reliable component of the rating mechanism. However, as AES systems are deployed in more areas in educational industry, it is of increasing importance to document how the scores they produce are related to the purpose of the assessment as a whole. By analyzing and improving upon the features these systems rely on to provide scores, the validity argument is strengthened in their favor.

AES systems developed to assess language proficiency particularly stand in need of this type of validation since AES systems built for other purposes are often tested on ELL essays to determine their reliability across domains (see e.g., Burstein & Chodorow, 1999; C.-F. E. Chen & Cheng, 2008; and Coniam, 2009). While some systems achieve reasonable results with this method, the validity argument suffers immensely.

5.1 Negation Acquisition in ELL Writing

In an attempt to boost the validity of AES systems used in ELL contexts, I plumbed the acquisition literature and a new ELL corpus for evidence of the developmental sequence of English negation in essays. The studies that had investigated and described such a sequence before all included small sample sizes that were conducted solely in the spoken domain. Thus, two aspects made my investigation unique: the magnitude of the data, which included over 1,700 students and over double that amount of essays, and the examination of this phenomenon in the written domain.

Dulay et al. (1982) provided the rationale to suspect that the acquisition order of negation as described in the spoken domain would not only surface in the essays, but that it would be mostly identical, as they suggested this had been the case with other acquisition sequences. Operating on this assumption, I extracted instances related to the characteristics described in the literature, proposing the slightly altered sequence in Figure 6.

Similar to what Housen (2002) found with the L2 acquisition of English verb forms, stage 1 of the negation sequence now seems to constitute “invariant default forms” that are applied to capture the semantic meaning of negation but which remain unanalyzed for tense, aspect or agreement. Stage 2 is where the majority of the overlapping is seen and where individual variation and L1 induced differences make any generalizations less potent. For the middle range of his composite sequence, Housen concluded that, “the variants behave like allomorphs as they appear first in random variation and then in complementary distribution. Their use is both underextended and overextended” (pg. 97). Finally, stage 3 is where consistent, successful use of the target-like forms separates those with higher proficiency from those at the lower levels. This is the stage at which there begins to be a functional analysis of forms in terms of their tense, aspect, and/or agreement morphology.

Previous research into interlanguage, to my knowledge, has not attempted to define where one acquisition stage ends and another begins in terms of the overall proficiency of the learner. Usually the analysis is of the acquisition order of a set of related structures relative to one another based on the productions of a group of subjects. The evidence I have collected provides at least some insight into how the acquisition of English negation relates to overall L2 English proficiency. Additionally, because this analysis involved a much larger sample size of ELLs, the reordered acquisition sequence proposed in this thesis could be a more accurate sequence than what has previously been documented.

One of the clear limitations encountered in this analysis was not having an equal set of essays at each scoring range. Because of this, additional mathematic means were called for in order to more clearly visualize the separation that indeed was there. Future work, however, could seek to corroborate these findings with an ELL written corpus with more even distribution or a large corpus of transcribed ELL speech to determine whether the order described here applies across domains or only to the written domain. Another study could follow a similar methodology as the present one and choose to focus on whether differences in the acquisition order exist based on the L1 of the learner as Pienemann et al. (1988) and Ahmad (2002) found for speakers of French and Spanish, respectively.

5.2 Using Features Related to Acquisition Order in an AES System

Based on the above findings, I encoded 31 derivative features to enhance a baseline scorer and tested it on the ELL corpus. The question of whether a scorer that includes these types of construct-relevant features is more accurate than one that does not is, in some ways, secondary to the goal of this thesis. However, while a set of features may work to bolster the validity argument, very few AES developers would ever include features that do not also show at least some improvement in the accuracy of the scorer (Vajjala, 2017). For this reason, the results of accuracy and agreement compared to human ratings were included and in each case the results for the enhanced scorer were slightly better than the baseline. Though modest, these trends also held for both types of training and testing splits tested.

Since I dealt with only one of the many possible acquisition sequences (i.e. negation), I did not expect drastically increased performance. As introduced previously, the text cohesion detector TAACO extracts over 150 indices thought to be related to discourse cohesion. Yet, with the ELL corpus, only 15 of the features were found to add ample improvement to the baseline scorer and were therefore implemented in the system. In the same light, Yannakoudakis & Briscoe (2012) experimented with 16 different sets of features related to discourse coherence and cohesion only finding three of them to modestly improve the Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient of their baseline scorer.

One of the drawbacks of the kind of research presented in this thesis that is common to most studies built on supervised machine learning techniques is that any number of models could be experimented with and individual parameters continuously optimized. Certainly the way in which I encoded the features of negation usage in this system could be adjusted, weighted, or otherwise optimized to further enhance their importance in the system. Until a standard is set for this type of research, the best option may be to adhere as closely as possible to default parameters in machine learning algorithms and encode the experimental features in an intelligible and replicable manner.

The improvements in accuracy in AES2 show that there is at least some basis for expecting that identifying and encoding more such phenomena in the future could lead to greater performance enhancers in AES scoring. Future research should encode other documented L2 acquisition sequences to see if they can improve the reliability of an AES system. A full set of indices related to the acquisition order of a variety of English structures would vastly improve the validity of AES systems developed for ELL essays. The preliminary analysis presented in this thesis should lead to research that involves analysis and testing with other learner corpora, other modalities (such as transcribed speech), other developmental sequences besides negation, and other L2 languages. Finally, as the use of AES expands to other languages, attested acquisition orders in those languages should be implemented into the systems built for those languages.

References

ACTFL. (2012). ACTFL proficiency guidelines [Electronic version]. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Ahmad, K. (2002). Don't just say "no": Developmental sequence of negation. *TESOL Working Papers Series*, *1*, 1–15.

Attali, Y. (2007). Construct Validity of e-rater® in Scoring TOEFL® Essays. *(ETS Research Report No. RR-07-21).* Princeton, NJ: Educational Testing Services.

Attali, Y. (2013). Validity and Reliability of Automated Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 181–198). Routledge.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2.0. *The Journal of Technology, Learning and Assessment*, *4*.

Briscoe, T., Medlock, B., & Andersen, Ø. (2010). *Automated assessment of ESOL free text examinations*. University of Cambridge, Computer Laboratory.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 68–75). Association for Computational Linguistics.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater Automated Essay Scoring System. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 55–67). Routledge.

Burt, M., & Dulay, H. (1978). Some Guidelines for the Assessment of Oral Language Proficiency and Dominance. *TESOL Quarterly*, *12*, 177–192.

Casillas, G. (2008). The insufficiency of three types of learning to explain language acquisition. *Lingua*, *118*, 636–641.

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*, 94–112.

Chen, H., & He, B. (2013). Automated Essay Scoring by Maximizing Human-Machine Agreement. In *Empirical Methods in Natural Language Processing* (pp. 1741–1752).

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays (TOEFL *Research Report No. 73; ETS RR-04-04)*. Princeton, NJ: ETS.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL*, *21*, 259-279.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237.

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

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17.

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*, 37–53.

Dulay, H., Burt, M., & Krashen, S. D. (1982). *Language Two*. New York: Oxford University Press.

Ellis, N. C. (2008). Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (Vol. 27, pp. 63–103). John Wiley & Sons.

Eskildsen, S. W. (2012). L2 Negation Constructions at Work. *Language Learning*, *62*, 335–372.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. *Language and Computers*, *52*, 123–145.

Hawkins, R. (2008). The nativist perspective on second language acquisition. *Lingua*, *118*, 465–477.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 77–116). John Benjamins Publishing Company.

Klima, E., & Bellugi, U. (1966). Syntactic regularities in the speech of children. In J. Lyons & R. Wales (Eds.), *Psycholinguistic papers* (pp. 183–208). Edinburgh, Scotland: Edinburgh University Press.

Krashen, S. D. (1985). *Second language acquisition and second language learning* (Reprinted). Oxford: Pergamon Press.

Lei, C.-U., Man, K. L., & Ting, T. O. (2014). Using learning analytics to analyze writing skills of students: A case study in a technological common core curriculum course. *IAENG International Journal of Computer Science*, *41*.

Lonsdale, D., & Strong-Krause, D. (2003). Automated rating of ESL essays. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2* (pp. 61–67). Association for Computational Linguistics.

Marcus, M. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank*. Fort Belvoir, VA: Defense Technical Information Center.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*, 57–86.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, *7*, 1–10.

Millett, R. (2006). *Holistic Scoring of ESL Essays Using Linguistic Maturity Attributes*. Master's thesis, Brigham Young University.

Milon, J. P. (1974). The Development of Negation in English by a Second Language Learner. *TESOL Quarterly*, *8*, 137–143.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*, 373–391.

O’Grady, W. (2008). The emergentist program. *Lingua*, *118*, 447–464.

Ortega, L. (2014). *Understanding Second Language Acquisition*. Routledge.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an Acquisition-Based Procedure for Second Language Assessment. *Studies in Second Language Acquisition*, *10*, 217–243.

Schumann, J. H. (1979). The acquisition of English negation by speakers of Spanish: A review of the literature. In *The acquisition and use of Spanish and English as first and second languages* (pp. 3–32). Mexico City.

Scott, J. (1999, January 31). Looking For the Tidy Mind, Alas. *The New York Times*. Retrieved from https://www.nytimes.com/1999/01/31/weekinreview/looking-for-the-tidy-mind-alas.html

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*, 209–232.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to Automated Essay Evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 1–15). Routledge.

Spada, N., & Lightbown, P. M. (1999). Instruction, First Language Influence, and Developmental Readiness in Second Language Acquisition. *The Modern Language Journal*, *83*, 1–22.

de Swart, H. (2009). Negation in early L2: a “window” on language genesis. In R. Botha & H. de Swart (Eds.), *Language Evolution: The View from Restricted Linguistic Systems* (Vol. 55, pp. 59–100). Utrecht: LOT.

Vajjala, S. (2017). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 1–27.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*, 85–99.

White, L. (2008). On the Nature of Interlanguage Representation: Universal Grammar in the Second Language. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (Vol. 27, pp. 18–42). John Wiley & Sons.

Williamson, D. M. (2009). A framework for implementing automated scoring. In *The annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).*

Williamson, D. M. (2013). Probable Cause: Developing Warrants for Automated Scoring of Essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 153–180). Routledge.

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

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 33–43). Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 180–189). Association for Computational Linguistics.

Yannakoudakis, H., & Cummins, R. (2015). Evaluating the performance of Automated Text Scoring systems (pp. 213–223). Association for Computational Linguistics.

Yi, B.-J., Lee, D.-G., & Rim, H.-C. (2015). The Effects of Feature Optimization on High-Dimensional Essay Data. *Mathematical Problems in Engineering*, *2015*, 1–12.

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 224–232).

Appendix A

Languages Represented in the Corpus

Language	**Essays**
Spanish	1594
Korean	620
Portuguese	447
Chinese	314
Japanese	202
Russian	92
French	59
Ukrainian	38
Mongolian	34
Thai	31
Haitian Creole	28
Italian	20
Arabic	19
Vietnamese	19
Creole	15
Armenian	13
Nepali	11
Bambara	9
Malagasy	7
Turkish	6
Bengali	5
German	4
Polish	3
Urdu	3
Farsi	3
Tunisian	3
Quechua	3
Mauritian Creole	2
Fulfulde	2
Hungarian	2
Bamanan	2
Tajik	2
Taiwanese	2
Aymara	2
Iranian	1
Albanian	1
Swedish	1
Amharic	1
Romanian	1
Belorussian	1
Persian	1
Unknown	9
Total	3632