Int J Artif Intell Educ (2018) 28:79–105
DOI 10.1007/s40593-017-0142-3

ARTICLE

Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features

Sowmya Vajjala[1]

Published online: 7 February 2017

Abstract Automatic essay scoring (AES) refers to the process of scoring free text responses to given prompts, considering human grader scores as the gold standard. Writing such essays is an essential component of many language and aptitude exams. Hence, AES became an active and established area of research, and there are many proprietary systems used in real life applications today. However, not much is known about which specific linguistic features are useful for prediction and how much of this is consistent across datasets. This article addresses that by exploring the role of various linguistic features in automatic essay scoring using two publicly available datasets of non-native English essays written in test taking scenarios. The linguistic properties are modeled by encoding lexical, syntactic, discourse and error types of learner language in the feature set. Predictive models are then developed using these features on both datasets and the most predictive features are compared. While the results show that the feature set used results in good predictive models with both datasets, the question ”what are the most predictive features?” has a different answer for each dataset.

Keywords Automated writing assessment · Essay scoring · Natural language processing · Text analysis · Linguistic features · Student modeling

Introduction

People learn a foreign language for several reasons such as living in a new country or studying in a foreign language. In many cases, they also take exams that certify their

✉ Sowmya Vajjala
sowmya@iastate.edu

[1] Iowa State University, Ames, IA 50011, USA

language proficiency based on some standardized scale. Automated Essay Scoring (AES) refers to the process of automatically predicting the grade for a free form essay written by a learner in response to some prompt. This is commonly viewed as one way to assess the writing proficiency of learners, typically non-native speakers of a language. Producing such an essay is also one of the components of many high stakes exams like GRE®, TOEFL® and GMAT®. Several AES systems are already being used in real world applications along with human graders. Along with such high stakes testing scenarios, AES could also be useful in placement testing at language teaching institutes, to suggest the appropriate level language class to a learner. Owing to this relevance to different language assessment scenarios, AES has been widely studied by educational technology researchers. While most of the published research on AES has been on proprietary systems, recent availability of publicly accessible learner corpora facilitated comparable and replicable research on second language (L2) proficiency assessment (Yannakoudakis et al. 2011; Nedungadi and Raj 2014).

One non-test taking application of automatic analysis of writing is in providing real-time feedback to the writers on their language use, highlighting their mistakes and suggesting ways for them to improve their writing. This has long been used as a part of word processing software (e.g., Microsoft Word), and more recently, tools such as grammarly.com, turnitin.com, e-rater® and WriteToLearn™ are being used in such scenarios for analysis of student writing. A related area of research, but not specific to language use, is automatic content assessment, typically for scoring short answers to questions (Burrows et al. 2015). Considering that many people enrolling in Massive Open Online Courses (MOOCs) are non-native English speakers, such automated methods could also be useful in providing feedback for them in scenarios where they have to write long essays as a part of evaluation for some of the courses, especially in arts and humanities. Thus, AES can also be useful in both these scenarios to assess the language used by the learners.

Despite this long history of research and real-life applications, very little is known about what linguistic features are good predictors of writing proficiency. One reason could be that most of the published research on the topic used only one data source to develop and evaluate their models. The generalizability of the linguistic features described in one approach to another dataset is also not explored much in existing research. In this background, this article explores the following research questions:

- Can we build good predictive models for automatically evaluating learner essays, based on a broad range of linguistic features?
- What features contribute the most to the accuracy of automatic essay scoring systems?
- Are the features generalizable to other datasets that have a different grading scheme and different target learner groups? That is, are the most predictive features the same across datasets, or is the predictive power of a feature specific to a corpus?
- What role does the native language of the writer play in their second language writing proficiency prediction?

These questions are addressed using two publicly available second language writing corpora containing text responses written in response to some prompt. Several

linguistic features that encode word, POS, syntax, discourse level information, and error features were developed using free, open-source NLP software, to build predictive models of L2 writing. While some of these features were used earlier in AES studies, some are newly introduced in this paper. The role of native language (L1) in L2 essay scoring has been documented in past research on AES (e.g., Chodorow and Burstein 2004) but has not been explored in great detail. This paper also takes a step in this direction and studies the role of L1 in predicting L2 proficiency. Thus, this research has a dual aim of understanding the linguistic properties of L2 writing proficiency (according to human graders) as well as developing predictive models for writing assessment. To summarize, the primary contributions of this paper are as follows:

- Predictive models of L2 writing for AES are constructed using a wide range of linguistic features, several of which have not been used in this context before.
- The question of the generalizability of linguistic features used for AES, which was not studied in detail earlier, is given importance, by exploring two datasets for this task. Existing research generally reported results only with a single dataset.
- The role of a learner's native language as a feature in predictive modeling of L2 writing proficiency was not explored much in past research. This paper takes a first step in this direction.

The rest of the paper is organized as follows: First, a collection of related work on this topic is discussed putting the current research in context. The two sections that follow describe the methodological framework in terms of the corpora used and the features studied. The next section describes the experimental setup, evaluation methods and results of the experiments. The paper concludes with a summary of the results, a discussion on the implications of the results and pointers to future work.

Related Work

Automated Essay Scoring (AES) has been an active area of research for about four decades now and several assessment systems such as Intelligent Essay Grader (Landauer et al. 2003), Project Essay Grade (Page 2003), E-Rater (Burstein 2003; Attali and Burstein 2006) and IntelliMetric (Elliot 2003) are being used in real-world applications as a companion to human graders. AES systems are typically developed using human scorer judgments as the gold standard training data to build automatic prediction models using textual features. Hence, the purpose of existing AES systems is to emulate a human grader. Contemporary AES systems use a wide range of features to build their predictive models, ranging from superficial measures like word length and sentence length to sophisticated natural language processing based approaches. The features employed to measure written language cover aspects such as: grammatical and orthographic correctness, language quality, lexical diversity and fluency (Dikli 2006). The definitions of individual aspects are typically based on human expert judgments. Some of the commonly used features in AES systems are: essay length, word length, counts of various parts of speech, syntactic structure,

discourse structure and errors. Distributional semantics based models such as Latent Semantic Analysis were also used in AES systems in the past (Landauer et al. 2003).

Despite a wide spread of AES systems, there is not a lot of published work on what are the different linguistic features that contribute to AES accuracy. One reason could be that the AES systems are proprietary software. Recent years saw the release of a few public datasets that can be used for developing AES models (e.g., Randall and Groom 2009; Yannakoudakis et al. 2011; Kaggle 2012; Blanchard et al. 2013). Yannakoudakis et al. (2011) released a corpus of learner essays written for Cambridge First Certificate in English (FCE) exam, and conducted several experiments to do predictive modeling of human scores on these essays. This was extended in Yannakoudakis and Briscoe (2012) who focused exclusively on how discourse coherence can be modeled for learner essays. More recently, Crossley et al. (2014) studied the relationship between linguistic properties and TOEFL scores using several linguistic indices based on Coh-Metrix (Graesser et al. 2012) and Writing Assessment Tool (Crossley et al. 2013b). They used a corpus of 480 essays scored on a scale of 1–5 for this study. In another study, Crossley et al. (2015) studied the role of various linguistic features in AES using a corpus of 997 persuasive essays written by students at four grade levels. In Crossley et al. (2016), they explored the relationship between cohesion features and writing quality using a corpus of 171 English learner texts. Apart from predictive modeling based approaches, there are also studies that focused on identifying distinctive linguistic features between proficiency levels (e.g., Tono 2000; Lu 2010, 2012; Vyatkina 2012) and L1 backgrounds (Lu and Ai 2015).

Most of the research in AES has been related to English writing owing to its widespread use and the availability of more learner corpora and language processing software for the language. However, the past half-decade saw the emergence of AES research in non-English (primarily European) languages. Ostling et al. (2013) developed an AES approach for detecting Swedish language proficiency using a corpus of high-school level exams conducted nationwide in Sweden. Hancke and Meurers (2013) described a proficiency classification approach for a publicly accessible dataset of German learner essays, based on the CEFR (Council of Europe 2001) scale used in Europe. Vajjala and Lõo (2013, 2014) developed an approach for automatically predicting Estonian learner proficiency on the CEFR scale, also based on a public dataset. In developing the features, all the above-mentioned approaches relied on the specific properties of the language (e.g., morphology) along with features generally used in English. However, to our knowledge, AES systems developed for non-English languages have not been put to widespread use in any real life application the way English AES systems are being used (yet).

Despite this background of a wide body of work, there is not much of published research on what specific linguistic properties of learner writing contribute to receiving a better proficiency score. Having such knowledge is not only useful for the improvement of AES systems, but also for systems that provide feedback to learners, such as e-rater® and WriteToLearn™. Exploring the relation between such linguistic features and proficiency scores is one of the issues addressed in this paper. While some of the features employed in this paper overlap with the features described in previous studies and all these papers essentially talk about the same underlying research

issues, the current paper differs in both the methodologies employed, and the overall features used. The second issue addressed in this paper relates to the generalizability of the linguistic features, which is addressed by multi-corpus evaluation using two large publicly available corpora.

There has been some recent research modeling task independence in AES (Zesch et al. 2015; Phandi et al. 2015), and they used another publicly available dataset of scored essays (Kaggle 2012). The research described in this paper differs from this strand of research primarily in two aspects: Firstly, they focus on modeling the differences between tasks, and adaptation of a model trained on one task data to another, and not on the specific features that contribute to the AES as such. Secondly, their experiments make use of only one corpus. Finally, one of the specific research questions in our case is to investigate native language influence on L2 proficiency prediction. This was not also modeled in these two articles as the native language background of learners in Kaggle dataset is not known.

Generalizability of a machine learning approach to essay scoring can be studied in two ways: multi-corpus study (training multiple models, each with a different training data) and a cross-corpus study (training with one corpus and testing it using another). We study feature generalizability using a multi-corpus study setup in this paper. Doing a multi-corpus study using two corpora that come from very different sources, and graded by human graders using different norms makes it possible for us to compare what features work for which corpus specifically. Since we do not know the exact guidelines for grading for individual corpora, it will not be appropriate to do a cross-corpus study in this scenario. Further, a cross-corpus study will only tell us about the generalizability of a model built using one dataset (with specific feature weights) on another. But a multi-corpus study will help us address our specific research question about feature generalizability i.e., what features are useful to develop predictive models across multiple data sources.

In terms of research on publicly available corpora, the current work can compare closely to Yannakoudakis et al. (2011) and Yannakoudakis and Briscoe (2012), who worked on the First Certificate of English corpus, which is one of the corpora used in this paper. In contrast with the pairwise-ranking approach used in their work, our model uses regression. While we model similar aspects of text as both these papers in our approach, the feature set described in this paper contains fewer, but denser features. It will be shown that our models achieve a comparable performance with the reported results on this dataset. Thus, compared to existing work on AES, this paper reports experiments with a new corpus, uses some new features that were not used in this context before, and compares the feature performance with more than one corpus. To our knowledge, this is the first multi-corpus study of automatic essay scoring task.

Methods: Corpora

The experiments were conducted using two publicly accessible corpora that have human scorer annotations of language proficiency for essays written by non-native English speakers. They are described below:

Table 1 TOEFL11SUBSET (1069 texts per category)

Proficiency	Avg. tokens	Avg. sentences
low	297.2	11.7
medium	378.3	16.1
high	421.9	17.8

TOEFL11 Corpus

Our first corpus for the experiments reported in this paper is the TOEFL11 corpus of non-native English (Blanchard et al. 2013). This is a collection of essays written by TOEFL iBT® test takers in 2006-2007 in response to the *independent writing* task in the test. The learners responded to one of the 8 prompts, each of which differed in the topic of the question and not in the task itself. The entire corpus consists of 12100 essays written by learners with 11 L1 backgrounds and belonging to three proficiencies (low, medium, high). These proficiency categories were a result of collapsed TOEFL scores that were originally on a scale of 1–6. The essays were written in response to eight prompts. The first version of this corpus that was released during the First Native Language Identification Shared Task (Blanchard et al. 2013) is used in this paper for the experiments. This version had essays sorted by the native language of the learners and had 900 texts per L1. However, the proficiency distribution was not even (5366 text samples for medium, 3464 for high, 1069 for low).

In the initial classification experiments, it was observed that this created a strong bias for the medium proficiency, which resulted in poor precision and recall for the other two proficiency categories. Hence, a sample of 1069 texts per category was selected, using the SpreadSubSample method implemented in WEKA toolkit (Hall et al. 2009) to train balanced classifiers described in later sections of this paper.[1] We will refer to this corpus as TOEFL11SUBSET for the rest of this paper. Table 1 shows average number of tokens and sentences per text for the three proficiency categories in our corpus.

This corpus has been primarily used for the native language identification shared task and its derivatives, and its usefulness for automated scoring of essays has not been explored much in research. The only available research that used this corpus for such a purpose is Horbach et al. (2015), and they rely exclusively on trigram frequency features, compared to the rich feature set described in this paper.

FCE Corpus

Our second corpus is the First Certificate of English (FCE) corpus that was publicly released by Yannakoudakis et al. (2011). They released a corpus of First Certificate of English exam transcripts, which is a subset of the larger Cambridge Learner Corpus (CLC). The corpus contains texts produced by takers of English as Second or Other

[1] The file ids of the text files used in this experiment can be shared for replication studies.

Language (ESOL) exams, and the associated scores on a scale of 1–40. These texts are typically letters and personal essays written in response to given prompts.

Since this corpus has a numeric scale, and with a much broader scale range than TOEFL11 corpus which had only three categories, we did not look into balancing the corpus for all the scores, as that would have resulted in a very small corpus rendering it useless for machine learning purposes. However, we used the same train-test setup as described in previous research that used this corpus (1141 texts from year 2000 for training, 97 texts from year 2001 for testing) for comparability. This will enable us to do a direct comparison with other reported research on this corpus. We will refer to these as FCE-TRAIN and FCE-TEST respectively for the rest of this paper.

It has to be noted that compared to TOEFL11 corpus where all prompts elicited same form of response (writing an analysis of agreement or disagreement with a given statement), the prompts in FCE corpus asked for different forms of response such as a letter, a report, a short story and so on.

As mentioned earlier, since the native language of the writer is one of the features considered in the models described below, Kaggle corpus (Kaggle 2012) was not considered in this study, as it does not provide this information about the writer's background. Additionally, Kaggle corpus also involved aggressive preprocessing for anonymization, with operations such as removing all named entities (Blanchard et al. 2013). This would also have affected some of our features, especially the discourse features.

Methods: Features

A broad range of features covering different aspects of linguistic modeling appropriate for learner texts were developed in this paper. There is no publicly available documentation on how exactly are the human scorers assigning scores to individual essays, apart from the general guidelines. Typically, AES research described features that can be automatically extracted, while encoding different linguistic aspects related to written language. The feature choices described in this article also follow a similar direction, while also considering what we know from Second Language Acquisition (SLA) about learner writing. L2 writing proficiency in SLA research has been discussed in the context of the notions of Complexity, Accuracy and Fluency (CAF) (Housen and Kuiken 2009). This paper describes automatically extractable features that can encode these three aspects at different levels of language processing. For complexity, features that study the lexical richness and syntactic complexity in SLA, and features from readability assessment research are used. For accuracy, features modeling errors made by learners are extracted. For fluency, features typically used to model discourse coherence of texts in computational linguistics were used.

In terms of the linguistic aspects encoded, the features are classified into 6 groups: word level, parts of speech, syntactic characteristics, discourse properties, errors, and others. All the underlying linguistic representations (tags, parses, coreference information etc.,) for the feature calculations described below are obtained from the

Stanford Parser (Socher et al. 2013) and the Stanford CoreNLP package (Manning et al. 2014).[2]

Word Level Features

This group of five features consists of the measures of lexical diversity typically used in first and second language acquisition studies as a measure of the diversity in a learner's vocabulary. While there are many measures of lexical diversity, we grouped some of the measures that do not rely on any linguistic representation other than words alone into this feature group. This group consists of four variations of type-token ratio (TTR) with the formulae described in Lu (2012), which are TTR ($num.types/num.tokens$), Corrected TTR ($num.types/\sqrt{(2 * num.tokens)}$), Root TTR ($num.types/\sqrt{(num.tokens)}$), and Bilog TTR ($Log\ num.types/Log\ num.tokens$) and a fifth measure called Measure of Textual Lexical Diversity (MTLD) as described in McCarthy and Jarvis (2010), which measures the average length of continuous text sequence that maintains the TTR above a threshold of 0.72. Since these features capture the diversity of language use, we can hypothesize that they will be useful for L2 proficiency prediction. While these features were used to do a corpus based analysis of their correlation with L2 proficiency (Lu 2012), they were not used in the context of predicting modeling of writing proficiency before, to our knowledge. This feature group will be referred to as WORDFEATURES in the rest of the paper.

Since all the five formulae primarily depend on the two variables -total number of tokens and total number of unique tokens in a text, there is a high degree of correlation between most of these **five** features. However, since our purpose is to develop predictive models and predictions need not necessarily be affected by such multicollinearity, all the features, including the correlated ones, are kept in the machine learning models. Feature selection will be performed in the later part of this paper, to select the most predictive features for models. It will also be shown later how a seemingly small change in the operationalization of the feature can change the direction of relationship.

POS Tag Density

This group consists of **27** features that are calculated based on the percentage of various POS tags in the text and some of the measures of lexical variation used in SLA research to assess learner texts Lu (2012). We took a subset of features described in this paper that rely on ratios between POS tags (Noun, verb, adjective, adverb and modifier variation formulas). All these POS tag based features will be referred to as POSFEATURES in the rest of the paper. They are listed in Table 2, grouped according to the motivations for these features. All features in this group, except the word variation features are calculated with the number of words in the text as

[2]The code for feature extraction is hosted at: https://bitbucket.org/nishkalavallabhi/ijaiedpapercode/.

Table 2 Features based on POS Tags

Lex. Variation features from Lu (2012)	General POS tags	Verb tags
POS_adjectiveVariation	POS_numNouns	POS_numVerbsVBD
POS_adverbVariation	POS_numProperNouns	POS_numVerbsVBG
POS_correctedVerbVariation1	POS_numPronouns	POS_numVerbsVBN
POS_modifierVariation	POS_numPerPronouns	POS_numVerbsVBP
POS_nounVar	POS_numAdjectives	POS_numVerbsVBZ
POS_squaredVerbVar1	POS_numAdverbs	POS_numModalVerbs
POS_verbVar1	POS_numConjunctions	
POS_verbVar2	POS_numInterjections	
POS_numLexicalWords	POS_numDeterminers	
	POS_numPrepositions	
	POS_numVerbs	
	POS_numWhPronouns	

the denominator. The variation features follow the formulae described in Lu (2012). While the lexical variation features have been used to study their correlations with L2 writing proficiency, they were not used in the context of predictive modeling before, to our knowledge. Some version of POS tag based features are used in almost all reported AES research, although the exact operationalizations may differ between this article and other research.

Syntactic Complexity

This group consists of 28 features extracted from the syntactic parse trees of sentences in learner essays. 14 of the features were used in the past to measure syntactic complexity in second language writing and its relation to writing proficiency (Lu 2010, 2011). These features are implemented based on the descriptions in Lu (2010) and using Tregex tree pattern matching tool (Levy and Andrew 2006) with syntactic parse trees, for extracting specific patterns. Although these 14 features were used to do a statistical analysis of second language writing in the past, they were not used in training any automated writing assessment models. The remaining 14 features estimate the average number and size of various phrasal groups (NP, VP, PP, WH-phrases) and measures of parse tree height (average height, number of subtrees, constituents, SBARs etc.) per sentence. This group will be referred to as SYNFEATURES in the rest of the paper. They are listed in Table 3. To our knowledge, these specific operationalizations of features (e.g., PPs per sentence, RRCs per sentence) have not been reported in AES research before.

Discourse Properties

Text coherence is one of the scoring criteria used in essay scoring (Burstein et al. 2010). Coherence also relates to the *fluency* aspect of the CAF framework in SLA

Table 3 Features based on syntactic parses

SLA features from Lu (2010)	Other Syntactic Features
SYN_avgSentenceLength	SYN_avgParseTreeHeightPerSen
SYN_MeanLengthofClauses	SYN_numSentences
SYN_MeanLengthofTunits	SYN_numConstitutentsPerSen
SYN_ComplexNominalsPerClause	SYN_numConjPPerSen
SYN_CNPerTunit	SYN_avgNPSize
SYN_ComplexTunitRatio	SYN_numNPsPerSen
SYN_CoordinatePhrasesPerClause	SYN_numPPSize
SYN_CoordPerTunit	SYN_numPPsPerSen
SYN_DependentClauseRatio	SYN_numRRCsPerSen
SYN_DependentClausesPerTunit	SYN_numSBARsPerSen
SYN_TunitComplexityRatio	SYN_numSubtreesPerSen
SYN_VPPerTunit	SYN_numVPSize
SYN_numTunitsPerSen	SYN_numVPsPerSen
SYN_numClausesPerSen	SYN_WhPhrasesPerSen

research, which makes it a relevant aspect to consider in L2 writing proficiency prediction models. However, there is no single method to model coherence and different approaches have been proposed in computational linguistics research so far. Hence, several discourse features are considered in this paper, that encode coherence with different levels of linguistic analysis, based on existing research in natural language processing.

At a very shallow level, there are a total of eight word overlap features - content word, noun, stem, and argument overlap at local (between adjacent sentences) and global (between any two sentences in a text) levels. The implementation of these overlap features is based on the descriptions in the Coh-Metrix tool (Graesser et al. 2012) documentation. This group of **eight** features will be referred to as DISC-OVERLAP in the rest of the paper. These features have been used in some of the recent writing assessment studies that used Coh-Metrix tool (Crossley et al. 2016).

At the level of part of speech, referential expressions like pronouns and articles are known to be some of the indicators of text cohesion. Hence, referential expression features based on the descriptions of Todirascu et al. (2013) (who used them to measure coherence and cohesion for readability assessment of texts) were implemented. This group consists of **10** features that model the use of definite articles and pronouns (all, personal, possessive) per word and per sentence, and pronouns and proper nouns per noun. These will be referred to as DISC-REFEX in the rest of the paper. To our knowledge, referential expression features were not explicitly studied in the context of automatic proficiency assessment before.

Discourse connectives (words such as: *and, although, but, however*, etc.,) that connect parts of a text are frequently used as a measure of text coherence in literature. Previous corpus research also emphasized the importance of studying connector usage in learner texts (Granger and Tyson 1996). Some of the existing tools like

Coh-Metrix[3] provide features that calculate the occurrence of various types of connectives in text based on word lists (e.g., causal, temporal, contrastive, etc.,). However, not all connective words are used always as discourse connectives in a text. Pitler and Nenkova (2009) described an approach to disambiguate the discourse versus non-discourse usage of connectives based on the syntactic parse trees. Along with this, they also provide a sense classification for the discourse usage of connectives (four senses: Expansion, Contingency, Comparison, Temporal). This method is available as a downloadable discourse connectives tagger,[4] which takes parse trees of sentences in a text as input, and tags it with discourse annotations. This tool was used to calculate **seven** connectives based features: the number of discourse and non-discourse connectives, all connectives, and the number of occurrences of each of the four senses, per sentence. This group will be referred to as DISC-CONN in this paper. To our knowledge, these features too have not been used for automatic proficiency assessment before.

Barzilay and Lapata (2008) introduced a model of text coherence based on the concept of an entity grid, where the various mentions of an entity in a text are labeled with their syntactic roles (subject, object, neither, and absent). These roles are then used to build a grid of entity transitions which capture how an entity changes across different sentences in a text. Burstein et al. (2010) and Yannakoudakis and Briscoe (2012) used entity grid based features for automated writing assessment before and they were among the more predictive feature groups in Yannakoudakis and Briscoe (2012). The current paper uses **16** features based on the transitions between the four syntactic roles for the entities (subject to subject, subject to object, subject to absent, subject to neither, and similar features for object, other, and neither respectively) and 4 additional features that calculate the average number of entities per sentence and per text, number of unique entities per text and average number of words per entity as additional features, as used in some of the earlier research in analyzing text coherence for readability assessment (Feng et al. 2010). This group of features is referred to as DISC-ENTITIES in this paper.

Co-reference chains refer to the chains of references to the same entity in a piece of text, as we progress from one sentence to another. They are a useful way to infer coherence in texts and were used in readability assessment research (Schejter 2015) as a measure of text coherence. In this paper, **8** co-reference chain features based on noun phrases, nouns and pronouns and determiner usage in the essay were used. These features will be referred to as DISC-CHAINS in this paper. To our knowledge, these features have not been used for AES earlier and this is the first article to report their utility for this task. Table 4 lists the coreference chain features.

Errors

Errors are one of the most intuitive features to use in assessing the proficiency of a writer. Some of the existing research on this topic used large language models with

[3] http://cohmetrix.com/.

[4] http://www.cis.upenn.edu/~nlp/software/discourse.html.

Table 4 Reference Chain Features

average length of a reference chain
proportion of personal pronouns in a reference chain
proportion of demonstrative pronouns in a reference chain
proportion of reflexive pronouns in a reference chain
proportion of proper nouns in a reference chain
proportion of possessive determiners in a reference chain
proportion of demonstrative determiners in a reference chain
proportion of indefinite Noun Phrases (NP) in a reference chain
proportion of definite NPs in a reference chain

word and POS unigrams to model learner errors (e.g., Yannakoudakis et al. 2011). In this paper, an open source rule based spelling and grammar check tool called LanguageTool[5] was used to calculate spelling and grammar features. The following **four** error features were extracted for each text: average number of spelling errors, non-spelling errors and all errors per sentence, and the percentage of spelling errors for all errors in a text. This group of features will be referred to as ERRORFEATURES in this paper.[6]

Document Length

Document length is one of the features used in all AES systems and is known to be a strong predictor of proficiency. Hence we included it as a feature, encoding it as the number of tokens in the document. This will be referred to as DOCLEN for the rest of this paper. It can be argued that document length may not be a useful feature to have in language exams, as they usually have a word limit. However, it has been used in all research and production AES systems, and is known to correlate strongly with proficiency and hence has been used in this paper as one of the features.

Others: Prompt and L1

Apart from all the above-mentioned features, prompt (the question for which the learners responded with answers) and the native language of the learner (L1) are considered as additional (categorical) features for both datasets.

Linguistic properties of writing prompts were also shown to influence the student writing (Crossley et al. 2013a). Since there is prompt information for both datasets, this was included in the feature set. However, it has to be noted that prompt was used only as a categorical feature without including any features that account for prompt specific vocabulary in the learner essays.

[5] https://languagetool.org/.

[6] The primary reason for choosing tool is the fact that it is under active development, and provides both spelling and grammar check in one place, eliminating the need to develop those modules for this study. We are not aware of any other such off-the-shelf library for spelling and grammar check.

While L1 was not directly considered as a feature in the L2 essay scoring approaches before, the possible influence of L1 on L2 proficiency prediction was discussed in previous research (Chodorow and Burstein 2004). Recently, Lu and Ai (2015) showed that there are significant differences in the syntactic complexity of L2 English writing produced by college level students with different L1 backgrounds. L1 specific differences in terms of linguistic complexity were also seen in the feature selection experiments with TOEFL11 DATASET in the recent past using readability assessment features (Vajjala 2015, Chapter 7). On a related note, two recent studies (Tetreault et al. 2012; Kyle et al. 2015) showed that L2 proficiency influenced the prediction accuracy of native language identification. In this background, we can hypothesize that L1 can be useful as a predictive feature for proficiency classification.

Both TOEFL11SUBSET and FCE have the L1 information for texts. So, this was used as a feature for training models with both datasets. Though all the L1s in TOEFL11SUBSET corpus had sufficient representation across all proficiencies, FCE corpus had a large imbalance in L1 representation across the score scale, as it had learners with more diverse L1 backgrounds compared to TOEFLSUBSET. Nevertheless, it was used as one of the features in the prediction models. However, we can hypothesize that this may not be as useful with FCE dataset.

Word and POS ngrams were typically used in other AES systems in the past to model learner language and compare it with "native" English. They were also used as a means to measure error rates in learner language (Yannakoudakis et al. 2011). However, they were not used in this paper for the following reasons:

- The features described above were all chosen to model clear linguistic properties, and are hence dense in nature. Ngram features are relatively sparse features, which can become difficult to interpret, compared to the dense features, in terms of what they are encoding linguistically.
- Ngram features can also be topic and vocabulary specific, and may not be very informative if we model across prompts, which differ in lexical choices.
- One major reason for using ngram features in previous research was to model errors in learner language compared to native English writing. However, in this paper, error features are designed based on LanguageTool, which internally models errors based on linguistic rules and ngrams.

Experiments and Results

The experiments reported in this paper primarily belong to two categories: development of predictive models, and a study of the most useful features for predictive models. The overall approach for conducting the experiments and interpreting the results is described below:

Approach

The scores in the two datasets used in this paper are designed differently. While the TOEFLSUBSET has three categories, the FCE dataset has a numeric score

ranging from 1–40. Hence, the following two approaches were adapted to enable a comparison between them:

1. Treat TOEFLSUBSET prediction as classification and FCE dataset as regression according to how the datasets are designed, and then compare what features are more predictive for both datasets, using a common feature selection approach which will work for both classification and regression methods.
2. Since both datasets are ordinal in nature, we can convert TOEFLSUBSET too into a numeric scale (low = 1, medium = 2, high = 3) and compare both datasets using a common prediction algorithm. While it is possible to create a discretized version of FCE too, that will not consider the fact that both datasets are ordinal in nature. Hence, that conversion is not performed in this paper.

WEKA toolkit (Hall et al. 2009) was used to train and test our prediction models. For TOEFLSUBSET, the classification models were evaluated in terms of classification accuracy in a 10 fold cross validation setting. For FCE, the performance of the models was assessed in terms of Pearson correlation and Mean Absolute Error (MAE) performance on FCE-TEST set. While 10 fold CV may give us a better estimate about the stability of the prediction model, there is some published research on FCE dataset with the exact train-test split as the one used in this paper. Thus, it is possible to do a direct comparison of the results on this dataset with existing research. So we focus on the results on test set for this corpus, and use 10 fold CV to compare between TOEFLSUBSET and FCE datasets. For training the models, multiple classification and regression algorithms were explored. Sequential Minimal Optimization (SMO), which is a support vector machine variant, worked well among these algorithms. Further, WEKA has both a classification and regression variant for this algorithm. So, SMO was used to train all the models described below, with normalized feature vectors and linear kernel.

Classification with TOEFL11SUBSET

As mentioned earlier, both classification and regression experiments were performed on this dataset. The first classification model with this corpus consisted of all the features described in the Features section. This model achieved a classification accuracy of 73.2 %. Table 5 shows the confusion matrix for the classifications. It can be seen that the largest confusion exists between low–medium and medium–high and hardly any overlap exists between low and high proficiencies. Consequently, medium proficiency is the most difficult to predict. This is not surprising, and is to be expected, if we remember that language proficiency is a continuum. Further, the confusion matrix works as an evidence that the classification model is actually learning the continuum of relationship between low/medium/high proficiencies.

Removing prompt and L1 as features resulted in less than 0.5 % decrease in classification accuracy. So, it looked as if there is no effect of these two features on the corpus at least when we use the full feature set. With this number as our comparison point for further experiments, we studied the impact of individual feature groups for this task. For each feature group, a small ablation test is reported as well, removing prompt and L1 as features. Table 6 shows the results for the feature groups. Random

Table 5 TOEFL11SUBSET Classification Summary

Classified as –>	Low	Medium	High
low	**881**	180	8
medium	157	**666**	246
high	4	264	**801**

baseline refers to the accuracy in the case where the classifier always picks only one class irrespective of the input. In our case, since all the three classes are represented equally, this becomes 33 %. Document length alone as a feature resulted in a 64.3 % classification accuracy, but created a skew towards low and high proficiencies, thereby affecting the precision and recall for medium proficiency.

Looking at the performance of feature groups, it can be noticed that the features that do not need much linguistic analysis (docLen, POS, Word) performed better than more linguistically demanding features. Despite this, the model with all features outperformed the next best feature group (POS) by 5 %. Error features performed the poorest as a stand alone group, achieving only 51 % accuracy. However, it should be remembered that the current approach only considers 4 features based on errors, and only considers two broad classes of errors (spelling and non-spelling). The classification accuracies were also in general lower for the discourse feature groups compared to other groups. While prompt seemed to have very little influence over classification accuracies, dropping L1 resulted in a huge drop in performance for all the discourse feature groups (marked in bold in Table 6). This drop is large as 12 % for DISCCONN features and around 5-10 % for other discourse feature groups. However, when

Table 6 TOEFL11SUBSET Feature Group Performance

Feature Group	Num. Features	Accuracy	Accuracy without prompt	Accuracy without prompt and L1
Random Baseline	–	33.0 %	33.0 %	33.0%
DOCLEN	1	66.3 %	66.3 %	64.3 %
WORD	5	67.4 %	68.0 %	66.9 %
POS	27	68.2 %	67.8 %	66.4 %
SYN	28	63.6 %	63.3 %	61.1 %
DISC-ALL	49	61.4 %	61.8 %	59.2 %
DISC-OVERLAP	8	56.8 %	56.8 %	**52.4 %**
DISC-REFEX	10	48.8 %	48.8 %	**42.0 %**
DISC-CONN	7	48.7 %	48.7 %	**36.0 %**
DISC-ENTITIES	16	49.0 %	49.0 %	**40.5 %**
DISC-CHAINS	8	48.7 %	48.7 %	**39.4 %**
ERROR	4	51.0 %	51.3 %	48.2 %
All Features	114 +2 (prompt, L1)	73.2 %	73.1 %	73.0 %

all the discourse features are combined, removing L1 feature has resulted in only a 2 % reduction in accuracy. Though we do not have any hypothesis about the reasons for this L1 influence on specific discourse features, this seems to indicate that there are differences in the usage of discourse markers by people with different L1 backgrounds, across the proficiency levels. This could be an interesting direction to explore in future.

Feature Selection To understand how many features we really need to reach the performance of all features put together, a range of feature selection methods provided in WEKA were explored - which used information gain, gain ratio, one R classifier, correlation and distance from other instances as the characteristics for feature selection. Twenty to thirty features were sufficient to reach the classification accuracy of 72-73 % with all the feature selection methods explored, which was the accuracy with all features put together. In all cases, this smaller subset included features from all the five major feature groups (word, pos, syntax, discourse, errors). Since a detailed comparison of top features from individual feature selection methods is beyond the scope of this article, we will focus on top-10 features using one method, ReliefF (Robnik-Sikonja and Kononenko 1997), which can be applied for both categorical and numeric class values.

Regression with TOEFL11SUBSET

TOEFL11SUBSET corpus is released as a categorical corpus. However, proficiency ratings are ordinal in nature and are not categories independent from each other. Further, to enable a comparison with FCE which has a numeric scale, TOEFLSUBSET was modeled as regression, using SMOReg algorithm. The model achieved a Pearson correlation of 0.8 and a Mean Absolute Error (MAE) of 0.4 with actual proficiency values. The Pearson correlation of different feature groups with proficiency is summarized in Table 7, which follows the same pattern as the classification model in terms of the feature group performance and the influence of L1 on discourse features (marked in bold in Table 7).

The ten most predictive features according to ReliefF feature selection method for this dataset are shown in Table 8, excluding prompt and native language. The algorithm selects attributes by repeatedly sampling instances in a dataset and comparing the value of the selected attribute for the nearest instances to the current instance.

Since ReliefF selects features independently, it does not capture the fact that some of the features can be correlated, as in the case of type token ratio variants and verb variation variants in Table 8. Despite that shortcoming, it can be observed from the table that there are features belonging to all the five major categories of features represented in this paper. Additionally, an important observation is: despite the fact that the top 10 features have features from all groups, they all rely on relatively shallow representations of language. The discourse features in this group are all word overlap features, which do not require a deep discourse analysis of the text. These 10 features together achieved a correlation of 0.76 in a 10 fold cross validation experiment, which is very close to what was achieved with the whole feature set. However, this

Table 7 TOEFL11SUBSET Feature Group Performance

Feature Group	Pearson Corr.	Corr. without prompt	Corr. prompt and L1
DOCLEN	0.67	0.67	0.65
WORD	0.70	0.69	0.67
POS	0.74	0.74	0.72
SYN	0.62	0.62	0.56
DISC-ALL	0.66	0.65	0.61
DISC-OVERLAP	0.59	0.59	0.49
DISC-REFEX	0.40	0.40	**0.24**
DISC-CONN	0.40	0.40	**0.16**
DISC-ENTITIES	0.40	0.40	**0.13**
DISC-CHAINS	0.40	0.40	**0.15**
ERROR	0.55	0.55	0.40
All Features	0.80	0.79	0.80

model still had features related to different aspects of language, though that need not necessarily mean a deeper modeling of linguistic structures.

While feature selection methods that choose individual best features and rank them is a good approach to follow, it has to be noted that the ReliefF selection method used above is independent of the learning algorithm, and is more of a dataset characteristic. What will give a better understanding of how the features work together in a model is a comparison of the features with highest positive and negative weights in the regression model. This can also facilitate a comparison with another model's feature weights. Table 9 below lists the five most discriminative features with positive weights and five with negative weights in the TOEFL11 regression model. It has to be noted that the weights refer to the feature weights in the context of the overall model, and not their individual weights independent of other features.

Table 8 TOEFL11SUBSET Most Predictive Features

Feature	Group
Document length	–
Corrected Type Token Ratio	WORD
Root Type Token Ratio	WORD
percentage of spelling errors	ERRORS
num. sentences per text	SYN
squared verb variation	POS
corrected verb variation	POS
global stem overlap	DISC
global argument overlap	DISC
global noun overlap	DISC

Table 9 TOEFL11SUBSET Feature Weights

High positive weight			High negative weight		
Weight	Feature	Feature group	Weight	Feature	Feature group
+1.31	document length	–	−1.51	num. non-spelling errors	ERRORS
+0.78	num. temporal connectives	DISC	−0.85	num. proper nouns	POS
+0.57	num. interjections	POS	−0.72	num. spelling errors	ERRORS
+0.49	pronouns to nouns ratio	DISC	−0.72	num. personal pronouns	DISC
+0.48	num. T-units per sentence	SYN	−0.60	num. non-discourse connectives	DISC

It is not surprising to see that two error features are among the features with high negative weights, which implies that a higher number of errors results in a low proficiency prediction. Longer texts result in a higher proficiency score, as is evidenced by document length feature having a high positive weight. POS and Discourse features are seen in both positively and negatively discriminating features, and there is a syntactic feature from SLA (num. T-units) among the most predictive positive weight features, which is consistent with the results from analysis of L2 writing in SLA research Lu (2010). Number of non-discourse connectives got a negative weight for this model, whereas number of temporal connectives got a positive weight. This implies that the sense of usage of connective words, and whether they are used as connectives at all or as normal words in the text plays a role in L2 writing proficiency assessment.

With FCE Corpus

Now, we turn to the question of how much these observations hold when we use a different corpus. Such an experiment will enable us to compare between the two datasets and observe what features work on both datasets and what work with only one dataset. To explore this direction, a regression model was trained with all the features using FCE-TRAIN as the training set in a 10-fold CV setting. The model achieved a Pearson correlation of 0.63 and a Mean Absolute Error (MAE) of 3.4 with the actual scores. On testing with FCE-TEST (containing 97 texts), the model had a correlation of 0.64 and a MAE of 3.6.

The same train-test setup was used in Yannakoudakis et al. (2011) and Yannakoudakis and Briscoe (2012). In the first study, they reported a highest correlation of 0.74 between the predicted and actual scores. In a second study with coherence features, they reported a highest correlation of 0.75 between actual and predicted scores. As we are testing on the same group of texts as previous research, it is possible to compare the correlations in terms of statistical significance. The comparison was performed using *cocor* correlation comparison library in R (Diedenhofen and Musch 2015) which compares two correlations using a variety of statistical tests and there was no significant difference between the correlations reported in Yannakoudakis

et al. (2011) and what was reported in this paper (0.64) with any of the tests. However, the result from the second study (Yannakoudakis and Briscoe 2012) is significantly better ($p < 0.05$) than the current model, which can perhaps be attributed to the presence of additional coherence features in their model (e.g., those based on incremental semantic analysis). Since these studies did not report a measure for error margins, we cannot compare the MAE values we got with existing work. Nevertheless, these results show that the feature set shows comparable performance with other reported results on this dataset.

Like with the experiments using TOEFL11SUBSET, the next step with this corpus too is to study which features contribute to a good prediction. Table 10 shows a summary of model performance with feature groups on the FCE-TEST data. For each feature group, we built three models - one with all features from the group, one excluding prompt and one excluding native language. However, in all the cases, removing prompt and native language did not have a significant effect on the results. So, only the results with all features (= features of the group+prompt+nl) are reported in the table.

Discourse features as a group seem to be the single best performing feature group for this data set, followed by POS and ERROR features. Within the discourse features, lexical chain features that require coreference resolution are the best performing features, while all other discourse features except referential expressions performed poorly on this corpus. In general, compared to the TOEFL11SUBSET, stand alone feature groups perform poorly compared to the entire feature set put together. Document length, which was very predictive of the performance for TOEFL11 dataset did not do particularly better in comparison with other features for FCE dataset.

To understand which individual features had the most predictive power on this dataset, we again used the ReliefF algorithm. Table 11 shows the ranked list of 10 best features and their categories.

Discourse features are the most represented group in the top 10 features, followed by word and error features. It is interesting to note that compared to the best

Table 10 Feature Group Performance on FCE-TEST

Feature Group	Pearson Correlation	MAE
DOCLEN	0.31	4.6
WORD	0.29	4.8
POS	0.49	4.4
SYN	0.44	4.4
DISC-ALL	0.55	4.0
DISC-OVERLAP	0.19	4.7
DISC-REFEX	0.34	4.8
DISC-CONN	0.04	5.0
DISC-ENTITIES	0.09	5.1
DISC-CHAINS	0.43	4.3
ERROR	0.46	4.0
ALL FEATURES	0.64	3.6

Table 11 Best Features on FCE-TRAIN dataset

Feature	Group
Average Length of a reference chain	DISC
All Errors (spelling and non-spelling)	ERROR
proportion of possessive determiners in a ref. chain	DISC
Non-spelling errors	ERROR
Corrected Type Token Ratio	Word
Root Type Token Ratio	Word
document length	
proportion of demonstrative pronouns in ref. chain	DISC
num. conjunctions/num. words	POS
Co-ordinate phrases per T-unit	SYN

features on TOEFL11SUBSET, there is an overlap only with respect to the lexical diversity features (TTR and CTTR). With respect to the error features, while percentage of spelling errors was the most predictive feature or TOEFL11SUBSET, non-spelling errors and all errors were the most predictive for FCE. Additionally, co-reference chain features, which did not figure in the most predictive features for TOEFL11SUBSET were among the best features for this corpus.

As with TOEFL model, Table 12 below shows the individual features that had the most positive and negative weights for this model.

An interesting aspect of this list of features is that, document length is a negative valued predictor, while sentence length is a positive predictor for FCE texts. Compared to TOEFL table (Table 9), there are a few differences regarding the importance of features in the model. There is no overlap among the positively weighted features of the two models, except for pronouns to nouns ratio, which is a discourse feature modeling referential expressions. Among the negatively weighted features, while spelling and non-spelling errors as separate features carried more weight in the TOEFL model, the error feature that combines both of them had more

Table 12 FCE Feature Weights

High positive weight			High negative weight		
Weight	Feature	Feature group	Weight	Feature	Feature group
+0.31	complex nominals per T-unit	SYN	−0.35	TTR	WORD
+0.27	num. pronouns per noun	DISC	−0.35	document length	–
+0.25	proper nouns to nouns ratio	DISC	−0.34	num spelling and non-spelling errors	ERRORS
+0.24	average sentence length	SYN	−0.32	num. proper nouns	POS
+0.22	corrected TTR	WORD	−0.28	T-unit complexity ratio	SYN

weight for the FCE model. Number of proper nouns is the feature that occurs commonly between TOEFL and FCE datasets with a negative weight. It is interesting to see the two variations of type-token ratio - TTR (types/tokens) and corrected TTR (types/$\sqrt{2 * tokens}$) appear on either side of the weight table for FCE dataset. Further, while proper nouns per sentence, POS feature, had a negative weight, proper nouns to nouns ratio, a referential discourse feature, had a positive weight in the model. Thus, the exact operationalization of the features can change the direction of the relationship with the score variable. This shows the usefulness of considering closely related features in the prediction model. It would not have been possible to understand these relations otherwise.

A common criticism of using a wide range of features in essay scoring has been that several features are merely proxies of document length. While this may be true to some extent, the results so far stress the importance of considering diverse features for constructing the predictive models. Further, the dependency on document length can also be an artifact of the data. One way to verify if that is the case is to look at the partial correlations of the features controlling for document length. This was done using *ppcor* R package (Kim 2015). For example, consider the feature corrected TTR. In the TOEFL corpus, the feature had a correlation of 0.6 with proficiency, but that dropped to a partial correlation of 0.3 when controlled for document length. However, in FCE corpus, the feature had a correlation of 0.36, and the partial correlation when controlling for document length was 0.25, which is not a drastic drop compared to TOEFL corpus. Similarly, consider another feature, *average length of the reference chain*, which is among the more predictive features of FCE corpus. One would expect that this feature, by definition, would highly vary with document length. That is, short documents cannot have very long reference chains owing to their length. So, a text with a long reference chain can be expected to have larger number of words. Thus, any correlation of this feature with proficiency can possibly be due to the document length. In TOEFL corpus, this feature had a very low correlation of -0.02 with proficiency, and the partial correlation controlling for document length is -0.13. However, with FCE corpus, this feature had a correlation of -0.43 with proficiency, but the partial correlation after controlling for document length was still -0.42. Thus, the dependence on document length of a certain feature could be an artifact of the corpus too, in some cases.

So far, from these experiments, the differences between the two datasets can be summarized as follows:

1. The two datasets differ from each other not only in terms of features that individually have a better predictive power with the essay score, but also in terms of features assigned high or low weights in a prediction model.
2. Features requiring deeper linguistic analysis (reference chain and syntactic complexity features) are more predictive in FCE dataset whereas the most predictive features of TOEFL11SUBSET were dominated by features requiring relatively shallow linguistic analysis, even when covering discourse aspects (e.g., word overlap).
3. While document length was the most predictive feature in TOEFL dataset, it had a relatively low correlation with essay score in FCE dataset.

4. Removing native language as a feature drastically reduced the classification accuracy for discourse feature groups in TOEFL11SUBSET. However this feature did not have any influence in the FCE dataset.

One reason for native language not having any effect in FCE could be due to the size of the dataset, in which there are learners with too many native language backgrounds, and each language does not have enough representation for a machine learning model to learn. Further, some languages from the training set are not seen in the test set at all. So, perhaps this dataset is not necessarily suitable to analyze L1 influence on L2. Other reason for having different features among the top predictors for both datasets may lie in the fact that they are texts of a different genre of writing. FCE texts contained more of narrative/conversational writings compared to TOEFL essays where the prompts always asked the writers to take a stance on some topic. The fact that FCE had a wider scale compared to TOEFL11 could be another factor. Further, since the grading rubrics used by both exams, and the process of grading are different, we can also infer that language proficiency means different things to different grading schemes. So, it would not be a wise idea to use a model trained on one dataset with another, in such a case. However, a comparison with another corpus will still be useful in providing such insights about the nature of language proficiency, while also providing us an idea of how to interpret the output prediction of a given AES system.

Thus, the primary conclusions from these experiments so far are:

- There may not be single best group of features that work for all datasets. One dataset may require features that demand deeper linguistic processing, and another dataset may not.
- Considering a larger group of features covering various linguistic aspects and levels of processing (instead of having a large group of features covering one aspect) however gave us better predictive models in both cases.
- The relationship of native language to L2 proficiency, especially with respect to the discourse features, needs to be explored further in future work.

Conclusions

This paper dealt with the issue of what linguistic features are more predictive for automatic scoring of English essays. To answer this question, several automatically extracted linguistic features that encode different aspects of language were explored, using free and open language processing software and resources. While some of these features were used for this task before, several of them were newly used in this paper. The usefulness of these features was studied by building predictive models using two public datasets - TOEFL11SUBSET and FCE. This makes the paper the first multi-corpus study for this task, and based on non-proprietary datasets. The paper's conclusions in terms of the original research questions from the start of the paper are as follows:

- *Can we build good predictive models for automatically evaluating learner essays, based on a broad range of linguistic features?*

TOEFL11SUBSET, the best model achieved a prediction accuracy of 73 % for classifying between three proficiencies (low, medium and high), using all the features. Almost all of the classification errors fell into differentiating between low–medium and medium–high with less than 1 % of the classification errors happening between low–high. When modeled as regression, the best correlation with proficiency of 0.8 was achieved with a model using all features. With FCE, the best model achieved a correlation of 0.64 and a Mean Absolute Error of 3.6, on the test data. This performance is comparable to other results reported on this corpus. In this backdrop, we can conclude that the feature set can result in good models for scoring the English essays of non-native speakers, with the currently available NLP approaches for pre-processing and feature extraction. It has to be noted, however, that these conclusions are restricted to what is possible with currently available language processing methods, and the predictive power of the features is also dependent on the accuracy of the underlying tools.

- *What features contribute the most to the accuracy of automatic essay scoring systems?*

  In TOEFL11SUBSET, document length was the best single predictor, and word and POS level features turned out to be the best feature groups. Looking at the features individually, there is at least one feature for each of our five major feature groups among the 10 most predictive features, and the features primarily relied on shallow linguistic representations of the language aspect. For example, shallow discourse features such as word overlap were more predictive for this dataset compared to deeper ones like reference chains. Error features were among those that got high negative weight in this model, and number of T-units per sentence, which is known in SLA research to correlate with language proficiency, was among the features with high positive weight. Removing native language as a feature resulted in a drop in classification accuracy for models using discourse features. So, native language of the author seems to be a useful feature to consider while modeling the discourse of L2 writing.

  In the FCE dataset, discourse features were the best performing group followed by POS and Error features. Native language and prompt did not have any effects on this dataset, and non-spelling errors were better predictors than spelling errors. While document length was among the best predictors, it was not as influential as in TOEFLSUBSET. In terms of the individual features, this dataset also had features from all five groups among the best features. However, features that relied on deeper linguistic modeling (such as reference chains) had more weight compared to other features. Error features and SLA based syntactic features were among the features with high positive or negative weight in this dataset too.

  One important aspect to remember is that the description of feature contributions here is limited to specifically features that are useful in the development of predictive models for scoring essays in terms of the language proficiency, and does not necessarily mean features that generally capture the broader construct of language proficiency, which involves aspects beyond essay writing.

- *Are the features generalizable to other datasets, used with a different grading scheme and different learner groups? Are the most predictive features same across the datasets, or is the predictive power of a feature specific to a corpus?*

  From the current experiments, the features do not seem to be completely generalizable across datasets. There is some overlap between most predictive features of both dataset, but not to the extent that we can claim generalizability. These results lead us to conclude that there may not be a universal feature set that predicts proficiency across all L2 writing datasets. But, the best models are always achieved by modeling multiple language dimensions in the features, and by considering deeper processing of texts. While we require features that encode multiple aspects of language from word level to discourse level, what features are more important depends on the nature of the dataset used. This could also directly relate to the fact that the two datasets originate from different exams, with probably different guidelines for grading and evaluation.
- *What role does the native language of the writer play in their second language writing?*

  Native language of the author as a feature was an important predictor for some feature groups in one dataset, but the effect was not seen in the other dataset. In the TOEFL11SUBSET experiment, there was a clear drop in performance of discourse feature based models when native language was removed from the feature set. While the effect was not seen in FCE dataset, it could also be because of insufficient representation of all native languages in the data. Nevertheless, the results on TOEFL dataset lead us to a conclusion that native language of the author can be a useful predictor for second language scoring models. Further analysis is needed in this direction to explore the influence of native language on the discourse properties of learner essays.

Future Work

One missing aspect in the models described here is that the actual word and phrase usage in the corpus has not been modeled at all in the features. While it is possible that the features currently considered indirectly capture such information, using features based on topical word/n-gram frequency measures, modeling prompt specificity explicitly should be pursued in future work. Another useful direction to take for handling this aspect will be to consider task independent features (e.g., Zesch et al. 2015 for modeling learner language.

Prompt has only been considered as a categorical feature in this paper, whereas it clearly needs to be better modeled in the context of recent research that showed how topic affects L2 writing quality (Yang et al. 2015). Modeling if the essay actually answers the question asked in the prompt is also an important aspect which needs to be addressed in future, which will also make AES a useful tool in scoring responses to assess the comprehension of learners. Adding the question relevance in these models will make it useful for large scale content assessment, in addition to assessing language form. The influence of native language on discourse features performance in the TOEFL11SUBSET is an interesting observation and needs to be investigated in better detail to understand the role of native language in learners'

written discourse. Looking at particular error types instead of two broad categories (spelling, non-spelling) as done in the current article may result in a better modeling of errors in L2 writing. Further, a qualitative analysis of the learner essays in terms of the most discriminative features and why they differ in terms of positive and negative weights with different datasets can give us better insights into the development of automated writing feedback systems. Finally, having a better understanding of the exact grading criteria used by human graders can result in a more theoretically grounded feature design in future.

Acknowledgments I would like to thank all the anonymous reviewers and the editors of the issue for their useful comments, which greatly helped improve this paper from its first version. I also thank Eyal Schejter for sharing his code to extract the coreference chain and entity density features.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: an entity-based approach. *Computational Linguistics*, *34*(1), 1–34.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). *TOEFL11: A Corpus of non-native english*. Princeton, New Jersey: Educational Testing Service. Technical Report ETS-RR-13-24.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.

Burstein, J. (2003). The e-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing, chapter 7, Lawrence Erlbaum Associates.

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays., *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 681–684). Association for Computational Linguistics: Los Angeles, California.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating E-Rater"s performance on TOEFL essays. *ETS Research Report Series*, *2004*(1), 1–38.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Crossley, S.A., Varner, L.K., & McNamara, D.S. (2013a). Cohesion-based prompt effects in argumentative writing., *FLAIRS Conference* (pp. 202–207).

Crossley, S.A., Roscoe, R.D., & McNamara, D.S. (2013b). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system., *FLAIRS 2013 - Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp. 208–213).

Crossley, S.A., Kyle, K., Allen, L.K., Guo, L., & McNamara, D.S. (2014). Linguistic micro features to predict l2 writing proficiency: a case study in automated writing evaluation. *The Journal of Writing Assessment 7*(1).

Crossley, S.A., Kyle, K., & McNamara, D.S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *The Journal of Writing Assessment, 8*(1).

Crossley, S.A., Kyle, K., & McNamara, D.S. (2016). The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1–16.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment (JTLA)*, *5*(1), 4–35.

Elliot, S. (2003). Intellimetric: From here to validity. *Automated essay scoring: A cross-disciplinary perspective*, 71–86.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment., *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (pp. 276–284). Beijing, China.

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2012). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

Granger, S., & Tyson, S. (1996). Connector usage in the english essay writing of native and non-native efl speakers of english. *World Englishes*, *15*(1), 17–27.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The weka data mining software: an upyear., *The SIGKDD Explorations*, (Vol. 11 pp. 10–18).

Hancke, J., & Meurers, D. (2013). Learner Corpus Research 2013. *Exploring CEFR classification for german based on rich linguistic modeling*. Bergen, Norway: Book of Abstracts.

Horbach, A., Poitz, J., & Palmer, A. (2015). Using shallow syntactic features to measure influences of l1 and proficiency level in efl writings., *4th workshop on NLP for Computer Assisted Language Learning* (pp. 21–34).

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

Kaggle (2012). The Hewlett Foundation: Automated Essay Scoring Competition. [Online; accessed 21-June-2016].

Kim, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, *22*(6), 665.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief., *Proceedings of the European Conference on Machine Learning* (pp. 171–182).

Kyle, K., Crossley, S.A., & Kim, Y.J. (2015). Native language identification and writing proficiency. *International Journal of Learner Corpus Research*, *1*(2), 187–209.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A cross-disciplinary perspective*, pp. 87–112.

Levy, R., & Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures., *5th International Conference on Language Resources and Evaluation* (pp. 2231–2234). Genoa, Italy.

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

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, *96*(2), 190–208.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level english writing: Differences among writers with diverse l1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit., *ACL (System Demonstrations)* (pp. 55–60).

McCarthy, P., & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

Nedungadi, P., & Raj, H. (2014). Unsupervised word sense disambiguation for automatic essay scoring., *Advanced Computing, Networking and Informatics-Volume 1* (pp. 437–443). Springer.

Ostling, R., Smolentzov, A., Tyrefors Hinnerich, B., & Höglin, E. (2013). Automated essay scoring for swedish., *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 42–47). Atlanta, Georgia: Association for Computational Linguistics.

Page, E.B. (2003). Project essay grade: PEG. *Automated essay scoring: A cross-disciplinary perspective*, pp. 43–54.

Phandi, P., Chai, K.M.A., & Ng, H.T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 431–439). Lisbon, Portugal: Association for Computational Linguistics.

Pitler, E., & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text., *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 13–16): Association for Computational Linguistics.

Randall, M., & Groom, N. (2009). The BUid Arab learner corpus: a resource for studying the acquisition of L2 english spelling., *Proceedings of the Corpus Linguistics Conference (CL)*. Liverpool, UK.

Robnik-Sikonja, M., & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression., *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 296–304).

Schejter, E. (2015). *Automatic Analysis of Coherence and Cohesion for the Assessment of Text Readability in English*. Germany: University of Tuebingen. Bachelor thesis.

Socher, R., Bauer, J., Manning, C.D., & Andrew, Y.N. (2013). Parsing with compositional vector grammars., *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 455–465). Sofia, Bulgaria: Association for Computational Linguistics.

Tetreault, J., Blanchard, D., Cahilll, A., & Chodorow, M. (2012). Native tongues, lost and found: Resources and empirical evaluations in native language identification., *Proceedings of COLING*, (Vol. 2012 pp. 2585–2602).

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L., & Bernhard, D. (2013). Coherence and cohesion for the assessment of text readability., *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science* (pp. 11–19).

Tono, Y. (2000). A corpus-based analysis of interlanguage development: analysing pos tag sequences of EFL learner corpora., *PALC'99: Practical Applications in Language Corpora* (pp. 323–340).

Vajjala, S. (2015). Analyzing text complexity and text simplification: Connecting linguistics processing and educational applications, University of Tübingen. Ph.D. thesis.

Vajjala, S., & Lõo, K. (2013). Role of morpho-syntactic features in Estonian proficiency classification., *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*: Association for Computational Linguistics.

Vajjala, S., & Lõo, K. (2014). Automatic cefr level prediction for Estonian learner text. *NEALT Proceedings Series*, *22*, 113–128.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, *96*(4), 576–598.

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., *Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 33–43).

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11* (pp. 180–189). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading., *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 224–232).