Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing

Scott A. Crossley°, Franklin Bradfield* & Analynn Bustamante°

° Georgia State University, Atlanta | USA

* Georgia Tech Research Institute, Atlanta | USA

Abstract: This study introduces GAMET, which was developed to help writing researchers examine the types and percentages of structural and mechanical errors in texts. GAMET is a desktop application that expands LanguageTool v3.2 through a user-friendly, graphic user interface that affords the automatic assessment of writing samples for structural and mechanical errors. GAMET is freely available, works on a variety of operating systems, affords document batch processing, and groups errors into a number of structural and mechanical error categories. This study also tests LanguageTool's validity using hand-coded assessment for accuracy and meaningfulness on first language (L1) and second language (L2) writing corpora. The study also examines how well LanguageTool replicates human coding of structural and mechanical errors in an L1 corpus as well as assesses associations between GAMET and human ratings of essay quality. Results indicate that LanguageTool can be used to successful locate errors within text. However, while the accuracy of LanguageTool is high, the recall of errors is low, especially in terms of punctuation errors. Nevertheless, the errors coded by LanguageTool show significant correlations with human ratings of writing and grammar and mechanics errors. Overall, the results indicate that while LanguageTool fails to flag a number of errors, the errors flagged provide an accurate profile of the structural and mechanical errors made by writers.

Keywords: grammar, mechanics, natural language processing, writing tool, assessment, writing quality

journal of
WRITING RESEARCH

Crossley, S.A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270. doi: 10.17239/jowr-2019.11.02.01

Contact:Scott A. Crossley, Department of Applied Linguistics/ESL, Georgia State University, 25 Park Place, Suite 1500, Atlanta, GA 30303 | USA - sacrossley@gmail.com

The ability to communicate effectively in writing is a strong predictor of academic and employment success (Graham & Perin, 2007). Writing effectively involves a number of processes including idea generation, argumentation, organization, rhetorical style, and the production and use of varied and sophisticated language features (Ferris, Eckstein, & DeHond, 2017; Flower & Hayes, 1981; Hyland, 2003; Zamel, 1982). In addition, effective writing depends on knowledge of the language system to include grammatical and syntactic rules (i.e., structural rules) along with knowledge of spelling and punctuation rules (i.e., mechanical rules). The importance of grammatical and structural errors can be seen clearly in writing assessment literature which highlights the importance of these errors in explaining writing quality (Eckes, 2008; Santos, 1988; Zhu, 2003), teaching these rules to students (Graham, 1983; Morris, Blanton, L. Blanton, W., & Perney, 1995), and teacher beliefs about the importance of these rules (Cutler & Graham, 2008). Additionally, structural and mechanical rules are important components of writing across a variety of disciplines and fields including business, engineering (Zhu, 2003), editing, and journalism (Ward & Seifert, 1990)

One problem with assessing the importance of structural and mechanical errors in writing is that coding such errors is time intensive, subjective, and prone to inaccuracies. While a number of systems are available that assess structural and mechanical in student writing such as Ginger, Grammarly, and Whitesmoke, these systems are meant to provide automated written corrective feedback to writers (Ranalli, Link, & Chukharev-Hudilainen, 2017) and are not designed for research purposes. In addition, the algorithms that inform these systems are not open-source and use of the systems are fee-based. Additional research on error detection focuses on automatic correction of errors in text (Ng et al., 2014). Thus, the need exists for a freely available automated structural and mechanical error analysis tool that is based on open-source algorithms that can be used by writing researchers to help better understand the importance of these errors in explaining writing proficiency - motivating writing instruction and furthering writing research.

The purpose of this study is to test the accuracy of LanguageTool v3.2 (Naber, 2003), which automatically groups language errors into a number of categories, including grammar, spelling, misspellings, typographical, white space, style, and duplication errors. We do this by developing a user-friendly interface for LanguageTool called the Grammar And Mechanics Error Tool (GAMET), which uses LanguageTool to automatically assess writing samples written in English for structural and mechanical errors and reports the errors counted in a usable data frame. GAMET works on a variety of operating systems, affords document batch processing, and is freely available at linguisticanalysistools.org. We test the precision of the features reported by Language Tool using hand-coded assessment for accuracy and meaningfulness on English first language (L1) and second language (L2) writing corpora collected from the intelligent tutoring system Writing-Pal (henceforth the Writing-Pal corpus) and the Test of English as a Foreign Language internet-Based Test (TOEFL-iBT, henceforth the TOEFL corpus) respectively. We also examine how well the LanguageTool replicates human coding of structural and mechanical errors in the Writing-Pal corpus (i.e., does it flag similar errors as human raters). Lastly, we assess correlations between GAMET and human ratings of essay quality for both the Writing-Pal and the TOEFL corpora.

1. Error Types

We follow the error type classification developed by Naber (2003) in which errors are categorized into four types: spelling errors, style errors, grammar/syntax errors and semantic errors. Spelling and grammar/syntax are likely the most common types of errors that writing researchers focus on. Spelling errors arise from not following conventional spelling rules in a language. Grammar/syntax errors are those errors in which sentence structure do not comply with a language's grammar or syntactic rules. Style errors comprise the use of uncommon words or phrases or the use of syntactic structures that make a text either too complicated or easy to understand. Style errors are context dependent and agreement on when a style error occurs may differ from person to person. Naber classifies all other errors that are not related to spelling, grammar, or style as semantic error. These types of errors require extensive background knowledge and are difficult if not impossible to detect automatically.

For this study, we focus mainly on spelling and punctuation errors (henceforth mechanical errors) and grammar/syntax errors (henceforth structural errors). The major difference between mechanical and structural errors is that punctuation and spelling are learned, while grammar is generally acquired implicitly during the acquisition of language. Thus, unlike grammar, spelling and punctuation require the use of formal instruction and serial memory (Henderson & Templeton, 1986) while grammar rules can be inferred by native speakers.

Research indicates that there is a strong link between mechanical errors and writing ability and quality (Morris, Blanton, Blanton, & Perney, 1995). For example, Wharton-McDonald, Pressley, and Hampston (1998) found that essays written by higher proficiency students were more accurate in terms of mechanics (Wharton-McDonald, Pressley, & Hampston, 1998). In terms of writing quality, Graham, et al. (1997) collected writing samples from 600 primary and intermediate grade children and analyzed the samples for compositional fluency and quality. Graham et al. found that mechanics accounted for a significant portion of the variance for both the fluency and quality scores. Crossley, Kyle, Allen, & McNamara (2014) identified relationships between essay quality and mechanical errors and found that spelling errors reported the largest correlations with human judgments of essay quality. They reported that a single spelling error index based on various spelling errors explained 9% of the variance in essay quality. These studies, along with others help to support the notion that mechanical aspects of writing are strongly related to writing quality potentially because mechanical errors may be related to interference of form over meaning (Graham, et al. 1997).

There is less evidence that structural errors are strongly related to writing quality and proficiency. For example, Crossley et al. (2014) examined a corpus of 100 persuasive essays that had been scored for holistic quality by expert raters. They reported on associations between these scores and hand-coded incidence scores of grammatical errors in the text. The results of the study found that grammatical errors in essays only had a small effect on expert judgments. Four grammatical features - article errors, verb morphology errors, noun errors, and verb errors - showed a correlation of at least .1 with essay scores, with none reporting a correlation greater than .15 leading Crossley et al. to conclude that grammatical errors were only weakly associated with writing quality.

2. Automated Error Analysis

A number of tools have been developed to assess grammar and spelling errors in texts, but these tools are generally used to provide feedback to writers or to automatically correct spelling in texts. The focus of the tools is not to provide writing researchers with information about structural and mechanical errors in texts. The earliest grammar tool was likely the Writer's Workbench (Macdonald, Frase, Gingrich & Keenan, 1982) which provided feedback to writers on a number of grammatical and spelling errors. The first widescale commercially available grammar and spelling tools was likely Criterion (Burstein, Chodorow, & Leacock, 2004), which was marketed by Educational Testing Services (ETS) and contains spelling and grammar error detection modules. Criterion is an automatic writing evaluation (AWE) tool that can identify errors in writing and provide feedback to high school and college users to increase successful text revisions. However, the algorithms that inform the models are not publicly available and the tool itself is fee-based, making access difficult.

Criterion has been empirically tested in a number of studies. As an example, Lavolette, Polio, and Kahng (2015) examined the errors reported by Criterion in a small sample of essays and found that 75% of the errors were accurate with the best accuracy rates reported for capitalization, missing comma, wrong word, and ill-formed verb errors (all of which reported accuracy rates over 85%. *Criterion* was least accurate with run-on sentences, incorrect article usage, and spelling errors, where reported accuracy rates reached lower than 50%. In a more recent study, Ranalli, Link, and Chukharev-Hudilainen (2017) found that *Criterion* performed above a baseline accuracy of 70% on four out of ten common errors scoring the highest accuracy rates for ill-formed verbs, subject-verb agreement errors, determiner-noun agreement errors and fragments. The system performed poorly on run-on sentences, incorrect word usage, preposition errors, missing or extra commas, and missing or extra articles. Users studies have compared errors between Criterion and expert instructors and English as a Second Language (ESL) writing. These studies demonstrate that Criterion fails to flag or mislabels many errors thought to be important to instructors (Dikli & Bleyle, 2014), as well as under-identified common ESL errors including pronoun and verb form errors (Ferris, 2011).

There have also been open-source approaches to developing error analysis algorithms and tools. A good example is the Computational Natural Language Learning (CoNLL) shared task on grammatical error correction (Ng et al., 2014) which aimed at the automatic identification and correction of grammatical errors in essays. Thus, unlike Criterion, identifying errors and providing feedback to users was not the focus of the CoNLL shared task. The winning entry in the 2013 task (Rozovskaya, 2013) targeted five grammatical mistakes common to ESL writers (article/determiner, preposition, noun number, subject-verb agreement, and verb form errors). The system used POS and shallow parsing along with language modeling for large native English corpora as input, achieving an F1 score of 0.31, where the F1 score is a balance between recall and precision of errors The winning system in the 2014 task applied a hybrid approach that included both rule-based error correction and statistical modeling to rank corrections, achieving a reported $F_{0.5}$ score of .373 (Felice et al., 2014) where $F_{0.5}$ is a combined measure of precision and recall, with double the weight given to precision over recall. The system performed best on adjective/adverb position errors, tone errors, and mechanical errors, while performing worst on re-ordering errors, run-on sentences, and wrong collocation/idiom errors.

3. Current Study

The purpose of this study is to introduce and provide validation for a structural and mechanical error analysis tool: the Grammar And Mechanics Error Tool (GAMET) for use by writing researchers examining texts written in English. GAMET provides a graphical user interface that allows researchers without background in computer science or NLP to automatically calculate the number and types of errors found in writing samples by providing a user-friendly interface for LanguageTool v3.2, allowing for greater accessibility. Thus, unlike Criterion, the purpose of GAMET is not to provide feedback to writers and unlike the CoNLL tasks GAMET will not automatically correct errors in writing.

The validation studies found in this study include human coding of errors flagged by GAMET to assess LanguageTool's classifications in terms of accuracy and meaningfulness in two corpora: the Writing-Pal corpus (i.e., an L1 corpus of persuasive essays) and TOEFL corpus (i.e., an L2 corpus of persuasive essays). For the Writing-Pal corpus, the errors reported by LanguageTool are compared to human coded structural and mechanical errors. In addition, the errors reported by LanguageTool are associated with human ratings of essay quality to examine the usefulness of GAMET in researching writing quality. For the TOEFL corpus, the errors reported by LanguageTool are only associated with human ratings of essay quality because hand-coded structural and mechanical errors were not available. The research questions that guide this study are:

1. Are the errors reported by GAMET accurate, meaningful, and related to writing quality in both the Writing-Pal and the TOEFL corpora?

2. Are the errors reported by GAMET similar to those coded by human raters in the Writing-Pal corpus?

4. Method

4.1 GAMET

The core functionality of GAMET is drawn from LanguageTool v3.2 (Naber, 2003). GAMET provides a GUI overlay to interact with LanguageTool and leverage its capabilities on one or more text documents. GAMET outputs its results to a .csv file that contains frequencies for individual errors as reported by LanguageTool as well as aggregated construct scores in which errors reported by LanguageTool are combined into similar constructs. GAMET also reports text length counts for the individual texts in order to compute normalized frequencies.

LanguageTool employs part-of-speech tagging and rule based-techniques to identify and classify errors. These techniques include hand-crafted and specific error patterns defined in XML documents as well as generic patterns defined as software routines to detect and classify errors. Benefits of this approach include the potential for crowd-sourced extension of the errors patterns it is capable of handling, as well as the ability to provide detailed feedback on flagged errors. LanguageTool first tokenizes a text at the sentence and word level, chunks sentences into phrases, and then assigns each word a POS tag. It then flags errors by matching token and chunk spans to pre-defined error patterns at the word, POS, and chunk level. In doing so, LanguageTool describes errors, but does not correct them. LanguageTool currently supports over 20 languages and allows users to add new XML-defined rules to target specific errors.

LanguageTool aggregates individual features to measure macro-errors in six broad categories: grammar, spelling, style, typography, white space and duplication, all of which are reported by GAMET. In addition, GAMET reports incidence counts for the micro-errors calculated by LanguageTool. For each error category, GAMET reports a raw total and a total normalized by text length. In practice, flagged micro-errors are rare and so the present study investigates the macro-error counts, which are described briefly below:

- ***Duplication errors.*** LanguageTool measures word duplications as errors (e.g., *you have have to know where you are*).
- ***Grammar errors.*** The grammar error index in LanguageTool measures word and phrase level errors. It includes verb errors such as errors with verb usage, person, tense and aspect as well as noun errors like pluralization or determiner agreement. It also reports on adjective errors (including comparative and superlative errors), adverb errors (including adverb word order), connector errors (including the incorrect uses of coordinators and subordinators), negation errors, and fragment errors.

- ***Spelling errors.*** The spelling error index in LanguageTool measures deviations from conventional dictionary spellings of words. It also includes lowercase errors, missing hyphens, and missing apostrophes in contractions.
- ***Style errors.*** Style errors reported by LanguageTool include features such as wordiness, redundancy, and word choice, among others.
- ***Typography errors.*** Typography errors in LanguageTool include not capitalizing when necessary, missing commas and possessive apostrophes, and other punctuation errors.
- ***White space errors.*** White space errors are calculated by LanguageTool when inappropriate spacing, either an unneeded space such as one before a punctuation or an instance where a space was needed but not inserted are found.

4.2 Testing Corpora

To test the accuracy of the features reported by the LanguageTool, we assessed the errors reported by GAMET on two writing corpora. The first corpus consisted of independent essays samples collected from two administrations of the TOEFL-iBT (i.e., the TOEFL corpus). The essays were composed by two groups of 240 test-takers who were stratified by quartiles for each task (N =480). The essays were written on two different prompts (one prompt per form). The essays, the final scores, and the demographic information of the test-takers were directly provided by the Educational Testing Service (ETS). The 480 test-takers included both English as a Second Language (ESL) and English as a Foreign Language (EFL) learners. They were from a variety of national and linguistic backgrounds.

Our second corpus comprised 100 essays from an on-line writing study conducted in the Writing Pal intelligent tutoring system that examined the potential for grammar and spelling errors to predict L1 writing quality (i.e., the Writing-Pal corpus; Crossley et al., 2014). The essays were written by public high school students in the metropolitan Phoenix area in the state of Arizona (in the United States). The students ranged in age from 14 to 19 and the majority of the students in the study were female (around 70%). About 60% of the students identified themselves as native speakers of English, with the remaining participants identifying themselves as non-native speakers of English but functionally bilingual. The essays were written on two prompts.

4.3 Human Ratings of Writing Quality

TOEFL essays. Two expert raters trained by ETS scored each essay using a standardized holistic rubric. The rubric describes five levels of writing performance, scored 1 through 5. In the rubric, linguistic sophistication at the lexical and syntactic levels is emphasized in addition to the development and the coherence of the arguments along with syntactic accuracy and lexical and grammatical errors. An independent essay with a score of 5 is defined as being a well-organized and developed response to the given topic, displaying linguistic sophistication and containing only minor language mistakes. In contrast, an essay with a score of 1 has serious problems in organization, idea development, language use, sentence structure or usage. The final holistic score of each essay was the average of the human rater scores if the two scores differed by less than two points. Otherwise, a third rater scored the essay, and the final score was the average of the two closest raters. While inter-rater reliability scores are not provided for the TOEFL-iBT scores in the public use dataset, Attali (2008) reported that weighted Kappas for similarly double scored TOEFL writing samples were .7.

Writing Pal essays. Two expert raters with at least 4 years of experience teaching freshman composition courses at a large university rated the quality of the essays using a standardized Scholastic Aptitude Test (SAT) rubric and an analytic rubric that contained four subsections: introduction, body, conclusion, and correctness (see Crossley et al., 2014 for more details). The correctness subsection consisted of one rating under the heading "grammar, syntax, and mechanics." The raters were asked to judge whether "The writer employs correct Standard American English, avoiding errors in grammar, syntax, and mechanics." Thus, individual ratings were not collected for grammar, syntax, and mechanics. The holistic score asked raters to judge essays based on developing points of view, critical thinking, use of examples and evidence, and text organization and focus. A higher rating on both indicated greater mastery.

Both the SAT and analytic rubric generated a rating with a minimum score of 1 and a maximum of 6. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. The final interrater reliability for all essays in the corpus was $r > .70$. The mean score between the raters was used as the final value for the quality of each essay. Additional information about the rating scale was reported in Crossley et al. (2014).

4.4 Human Error Codings

We used error codings reported in two previous studies for the Writing Pal essays. In the first study (Crossley & McNamara, 2011), each essay was scored by two trained raters for a grammar, syntax, and mechanics category in an analytic rubric. The rubric was based on a 1-6 scale and asked raters to judge whether "The writer employs correct Standard American English." A score of 1 would indicate the use of incorrect Standard American English and a score of 6 would indicate the use of correct Standard American English. The inter-rater reliability for the grammar, syntax, and mechanics ratings was $r = .740$

In the second study (Crossley et al., 2014), an error tag-coding scheme was developed to investigate the frequency of grammar, mechanics, word use, and spelling errors in the 100 Writing Pal essays. The coding scheme was based on an error tagging manual reported in Dagneaux, Dennes, Granger, and Meunier (1996). The manual consists of subsections related to form (spelling and morphology), grammar (nouns, adjectives, and verbs), lexico-grammar (complementation, dependent prepositions), lexical choices (single, phrases, connectors, and conjunctions), and word problems (redundant and missing words). Two expert raters were trained on this manual and new codes based on the errors found in the corpus were added to the coding scheme. The added errors related to punctuation, spelling, sentence fragments, and ambiguous referents. After training was completed, the raters coded each essay independently and codes between raters were compared. Differences in coding were adjudicated between the two raters until agreement was reached. Additional information about the human error ratings can be found in Crossley et al. (2014).

4.5 Human Coding of LanguageTool Output

Two expert raters were trained to classify all errors reported by LanguageTool in terms of accuracy and meaningfulness. The raters were undergraduate linguistic majors at a large southeastern university in the United States. The raters were first jointly trained on a subset of errors and later analyzed all errors independently. In terms of accuracy, the raters examined each error and scored it on a three-point scale based on whether the classification was not accurate, potentially accurate, or accurate. In terms of meaningfulness, the raters were trained to identify whether the errors identified by LanguageTool would interfere with the processing of the text (in terms of both text reading speed and in text comprehension) for expert raters and, as a result, influence expert ratings of writing quality.

A similar three-point scale was used to identify errors as not meaningful, potentially meaningful, or meaningful. Meaningfulness in the context of these ratings was related to prescriptive and descriptive language rules. Prescriptive rules are operationalized as language rules that are argued to be correct even though they are not strongly practiced by the larger community of speakers and writers. The language used by the community and, importantly, its structure, comprises descriptive rules. Meaningfulness was strongly related to whether the errors reported by LanguageTool were descriptive or prescriptive with a lower meaningfulness scores given to prescriptive errors. For example, LanguageTool flagged the use of "thrives" as an error in the sentence "The media thrives on it" because prescriptively "media" is plural and would not require the use of the third person inflectional morpheme found in "thrives." However, it is common for speakers and writers to use a third person inflectional morpheme with "media." Thus, the raters scored the error as accurate, but not meaningful because it would likely not have interfered with the processing of the sentence meaning.

After initial ratings, disagreements were adjudicated between the two raters where possible. Flagged errors that were scored as potentially accurate or meaningful were also adjudicated. Exact agreement for accuracy among the two raters after adjudication was high for the TOEFL essays (.965) and for the Writing Pal essays (.990). Agreement for meaningfulness after adjudication was strong for the TOEFL essays (.962) and for the Writing Pal essays (.999). When disagreements remained, a third, expert-rater adjudicated the disagreement. When agreement could not be reached, the error flagged by LanguageTool was classified as "uncertain" (i.e., when the raters could not agree if the flagged error was an actual error or not). Average scores for accuracy and meaningfulness across the corpora were used as dependent variables in statistical analyses.

4.6 Analyses

We conducted a number of descriptive and statistical analyses of the data to help validate LanguageTool and to assess links between the errors reported by LanguageTool and human ratings of writing quality. Descriptively, we provide information about the number and types of errors flagged by LanguageTool in both the TOEFL and the Writing-Pal corpora. We then examine the accuracy and the meaningfulness of the errors reported by LanguageTool as judged by the human raters. We also provide the number and percentage of "uncertain classifications". For the Writing-Pal corpus, we also provide counts and percentages of the errors flagged by the human raters and not flagged by LanguageTool and the meaningfulness of those errors as judged by human raters. Lastly, we examine correlations between holistic scores of essay quality for both the TOEFL and Writing-Pal corpus and the overall error counts and the categorical errors counts reported by LanguageTool.

5. Results

5.1 Number and Types of Errors

TOEFL Corpus. GAMET identified 7,087 errors in the 480 essays that comprised the TOEFL corpus. Of these errors, the majority were spelling errors followed by white space, grammar, typographical, uncategorized, duplication, and style errors (see Table 1 for details). The expert ratings of LanguageTool's accuracy indicated that LanguageTool was most accurate at identifying white space and spelling errors. It was least successful at identifying style errors (see Table 2 for accuracies). The majority of uncertain classifications (i.e., errors identified by LanguageTool that may or may not be errors according to the raters) were style errors.

Table 1. Types of errors identified by LanguageTool (TOEFL corpus)

Errors	Count	Percentage
Grammar	627	8.8
Misspellings	4154	58.6
Style	99	1.4
Typographical	454	6.4
Uncategorized	328	4.6
White space	1389	19.6
Duplication	37	.5

Table 2. Accuracy and meaningfulness of errors identified by LanguageTool (TOEFL corpus)

Errors	Classification errors (%)	Accurate classifications (%)	Uncertain accuracy classifications (%)	Not meaningful errors (%)	Meaningful errors (%)	Uncertain meaningful ness classificatio ns (%)
Grammar	50 (8)	501 (79.9)	76 (12.1)	69 (11.)	550 (87.7)	8 (1.3)
Misspellings	63 (1.5)	4048 (97.4)	43 (1)	78 (1.9)	4070 (98)	6 (.1)
Style	0 (0)	28 (28.3)	71 (71.7)	98 (99)	1 (1)	0 (0)
Typographical	18 (4)	424 (93.4)	12 (2.6)	23 (5.1)	428 (94.3)	3 (.7)
Uncategorized	14 (4.3)	310 (94.5)	4 (1.2)	288 (87.8)	34 (10.4)	6 (1.8)
White space	3 (.2)	1384 (99.6)	2 (.1)	3 (.2)	1385 (99.7)	1 (.1)
Duplication	0 (0)	35 (94.6)	2 (05.4)	0 (0)	36 (97.3)	1 (2.7)

Writing Pal Corpus. LanguageTool identified 1,036 errors in the 100 essays that comprised the Writing Pal essay corpus. Like the TOEFL corpus, the majority of the errors were spelling errors followed by white space, grammar, typographical, uncategorized, style, and duplication errors (see Table 3 for details). The expert ratings of LanguageTool's accuracy indicated that LanguageTool was most accurate at identifying white space and misspelling errors. It was least successful at identifying style errors (see Table 4 for accuracies). The majority of uncertain classifications (i.e., errors identified by LanguageTool that may or may not be errors according to the raters) were style errors.

Table 3. Types of errors identified by LanguageTool (Writing-Pal corpus)

Errors	Number	Percentage
Grammar	64	6.2
Misspellings	699	67.5
Style	34	3.3
Typographical	35	3.4
Uncategorized	187	18.1
White space	8	.8
Duplication	9	.9

Table 4. Accuracy and meaningfulness of errors identified by LanguageTool (Writing-Pal corpus)

Errors	Classification errors (%)	Accurate classifications (%)	Uncertain accuracy classifications (%)	Not meaningful errors (%)	Meaningful errors (%)	Uncertain accuracy classification (%)
Grammar	16 (25.)	34 (53.1)	14 (21.9)	19 (29.7)	45 (70.3)	0
Misspellings	74 (10.6)	620 (88.7)	5 (.7)	83 (11.9)	616 (88.1)	0
Style	19 (55.9)	15 (44.1)	0 (0)	34 (1)	0 (0)	0
Typographical	2 (5.7)	31 (88.6)	2 (5.7)	2 (5.7)	33 (94.3)	0
Uncategorized	2 (1.1)	185 (98.9)	0 (0)	187 (1)	0 (0)	0
White space	4 (50.)	4 (50)	0 (0)	4 (50)	4 (50)	0
Duplication	0 (0)	9 (1)	0 (0)	0 (0)	9 (1)	0

5.2 Meaningfulness of Errors

TOEFL Corpus. The expert ratings for the meaningfulness of the errors in the TOEFL essays identified by LanguageTool indicated that the majority of grammar, misspelling, typographical, white space, and duplication errors were meaningful. In contrast, errors related to style and uncategorized errors were not meaningful (see Table 2 for accuracies). There were fewer errors that were identified as uncertain classifications that were not classified as meaningful or not meaningful. In most cases, uncertain classification accounted for about 1% or less of all the errors rated.

Writing Pal Corpus. The expert ratings for the meaningfulness of the errors identified by LanguageTool for the Writing Pal essays indicated that the majority of grammar, misspelling, typographical, and duplication errors were meaningful (similar to that reported for the TOEFL error analysis). In contrast, all errors related to style and uncategorized errors were not meaningful (see Table 4 for accuracies). All errors that were identified as uncertain classifications were classified as either meaningful or not meaningful.

5.3 Comparison with Human Error Analysis

Of the 1,036 errors identified by LanguageTool in the Writing-Pal corpus, 918 of the errors were flagged as accurate by the human raters examining the LanguageTool output (accuracy = .886). Of the 1,036 errors flagged by LanguageTool, 585 had been identified by the human raters who initially coded the corpus for errors, giving GAMET a precision of 57%. Four of the errors were flagged by the initial raters, but were not judged to be accurate by the subsequent raters. Of the remaining 447 errors flagged by LanguageTool but not flagged by the initial human raters,

114 of the errors were judged to not be accurate. For the remaining 333 errors that were flagged by LanguageTool, not flagged by the initial raters, and judged to be accurate by the subsequent raters, 124 were judged to be meaningful while the remaining errors were judged to not be meaningful,

Within the corpus of errors, LanguageTool failed to flag 1,872 errors identified by the human raters, indicating a recall of .239 (overall F1 of .336). The majority of these errors were related to punctuation (see Table 5). Other errors were related to form, grammar, and word characteristics (missing words, words out of order, or redundant words). In terms of meaningfulness, the raters indicated that of the 1,872 errors that LanguageTool missed, 1,728 of the errors were meaningful (92%, see Table 6).

Table 5. Types and counts of errors missed by LanguageTool (Writing-Pal corpus)

Error type	Count	Percentage
Ambiguity	30	1.6
Form (spelling and morphology)	222	11.9
Fragments	40	2.1
Grammar	261	13.9
Lexico-grammar	73	3.9
Lexis	84	4.5
Punctuation	917	49
Words (missing, order, redundant)	245	13.1

Table 6. Meaningfulness of errors missed by LanguageTool (Writing-Pal corpus)

Error type	Not meaningful errors (%)	Meaningful errors (%)	Uncertain classifications (%)
Ambiguity	1 (3.3)	28 (93.3)	1 (3.3)
Form (spelling and morphology)	8 (3.6)	213 (95.9)	1 (.5)
Fragments	2 (5)	38 (95.0)	0 (0)
Grammar	15 (5.7)	241 (91.3)	5 (1.9)
Lexico-grammar	0 (0)	72 (98.6)	1 (1.3)
Lexis	4 (4.8)	77 (91.7)	3 (3.5)
Punctuation	65 (7.1)	833 (90.8)	19 (2.1)
Words (missing, order, redundant)	14 (5.7)	226 (92.2)	5 (2)

5.4 Correlations with Human Scores: TOEFL Corpus

Correlations were calculated between the TOEFL writing quality scores and the output from LanguageTool. The analyses demonstrated that LanguageTool's output in all seven error categories demonstrated at least a significant and small effect size (*r* > .10) with the human ratings. The correlations were negative, indicating that fewer errors led to higher scores of grammar and mechanical ability. The *r* and *p* values for these correlations are presented in Table 7.

Table 7. Correlation between holistic writing scores and LanguageTool error counts

Corpus	TOEFL		Writing-Pal	
Feature	*r*	*p*	*r*	*p*
Overall errors	-0.528	< .001	-0.203	< .001
Duplication errors	-0.146	< .001	-0.149	0.139
Grammar errors	-0.399	< .001	-0.139	0.168
Misspellings	-0.477	< .001	-0.183	0.068
Style errors	0.100	< .050	-0.030	0.764
Typographical errors	-0.254	< .001	-0.073	0.471
White space errors	-0.282	< .001	-0.042	0.676

5.5 Correlations with Human Scores: Writing Pal Corpus

Holistic scores. Correlations were calculated between the human scores for essay quality and the output from the LanguageTool. The analyses demonstrated that output in four of the LanguageTool's error categories demonstrated at least a small effect size (*r* > .10) with the expert ratings. However, because the sample size was small, only one variable demonstrated a significant relationship. The correlations were negative indicating that fewer errors led to higher essays scores. The *r* and *p* values for these correlations are presented in Table 7.

Grammar and mechanical scores. Correlations were calculated between the hand-coded scores for grammar and mechanical errors and the output from LanguageTool. The analyses demonstrated that LanguageTool's output in all seven error categories demonstrated at least a significant and small effect size (*r* > .10) with the human ratings. However, because the sample size was small, output in only four categories demonstrated significant *p* values. The correlations were negative indicating that fewer errors led to higher scores of grammar and mechanical ability. The *r* and *p* values for these correlations are presented in Table 8.

Table 8. Correlation between expert grammar, mechanic, and syntax scores and LanguageTool error counts (Writing-pal Corpus)

Feature	*r*	*p*
Overall errors	-0.514	< .001
Duplication errors	0.119	0.238
Grammar errors	-0.142	0.160
Misspellings	-0.510	< .001
Style errors	-0.107	0.288
Typographical errors	-0.203	< .050
White space errors	-0.233	< .050

6. Discussion

In this paper we have introduced GAMET, which is based on LanguageTool v3.2 (Naber, 2003), and provided validation metrics on the output of the LanguageTool. GAMET is meant to be used by writing researchers without a background in computer scientists and thus cannot easily access the metrics reported by LanguageTool. We envision that these researchers will use GAMET to assess the types and number of structural and mechanical errors found in student, professional, and personal writing. Thus, the purpose of GAMET is distinct from Criterion, which focuses on providing feedback to users and it is unlike the models reported for the CoNLL tasks, which focused on error correction. The accuracy of the errors reported by GAMET along with their meaningfulness indicates that LanguageTool features can be used to successful locate errors within text. However, while the precision of the errors reported by LanguageTool is high, the recall, at least for the Writing-Pal corpus, was low (i.e., the errors it locates are generally accurate, but it fails to detect many errors). The majority of the errors missed (49%) were punctuation errors while around ~10% of form, grammar, and word errors were missed. Nevertheless, the errors reported by LanguageTool show strong associations with human ratings of L2 writing and overall grammar and mechanics errors, indicating that while LanguageTool may fail to flag a number of errors, the errors flagged by LanguageTool do provide an accurate profile of the structural and mechanical errors made by writers.

In terms of accuracy, features reported by LanguageTool demonstrated accurate classifications (above 90%) for the majority of errors reported in the TOEFL corpus including misspellings, typographical errors, white space errors, and duplication errors. The accuracy was lower for grammatical errors (~80%). While it is difficult to compare accuracy rates across tools because reported studies differ in terms of task, topic, scoring metrics, training of raters, domains, and a variety of other features, it is important to note that the accuracy rates for spelling reported by LanguageTool are higher than previously reported for off-the-shelf spell checkers tested on L2 writing samples (e.g., the 62% accuracy reported for Microsoft Word spell checker by Rimrott & Heift, 2008). The accuracy rates for grammar, typographical, white space, and duplication errors are also higher than those reported by similar tools. For instance, Lavolette et al., (2015) reported that 75% of errors reported by *Criterion* were accurate. In a more recent study, Ranalli et al. (2017) found that *Criterion* performed above a baseline accuracy of 70% on four out of ten common errors. Again, while direct comparisons are difficult to make, the accuracy rate of errors flagged by LanguageTool seem on par or slightly better than those reported in previous studies.

The accuracy rates reported by LanguageTool decreased for writers in the Writing-Pal corpus. For these writers, high accuracies were reported for misspellings, typographical and duplication errors, but lower accuracies (~50%) were reported for grammar and white space errors. The style errors reported by LanguageTool were generally inaccurate for all corpora with classifications between 29% (L2 essays) and 44% (L1 essays). The lower accuracies for errors reported for the essays in the Writing-Pal corpus (when compared to L2 essays) may have resulted from L2 writers producing more common errors that are easier to distinguish computationally or because the errors they produce are obviously misstructured because they are based on translation from a first language, or because they are producing unique grammatical errors (Myles, 2002; Richards, 2015; Wei, 2015). In contrast, the L1 and functionally bilingual writers in the Writing-Pal corpus likely have developed strong enough language structural and spelling rules to avoid common errors. It may also be a matter of sample size in that the L1 and functionally bilingual sample in this study was much smaller than the L2 sample. Knowing that L1 and functionally bilingual writers should, on average, produce fewer errors, it is likely that a larger sample size may be needed to fully tag the type and diversity of L1 and functionally bilingual errors.

In terms of the meaningfulness of the errors reported (i.e., the degree to which the errors would affect processing of the text for expert raters), high accuracies between 88% and 100% were reported for grammar errors, misspellings, typographical errors, white space errors, and reduplication errors in the TOEFL corpus. Lower error meaningfulness was reported for the Writing-Pal corpus with meaningfulness at 50% for white space errors and 70% for grammar errors. Error meaningfulness was higher (88%-100%) for misspellings, typographical errors, and duplication errors. Like the accuracy ratings, the flagged errors in the Writing-Pal corpus may be less meaningful because they are less common.

While LanguageTool achieved high accuracy and meaningfulness metrics, its recall metrics (i.e., the number of errors it did not flag) were low for the Writing-Pal corpus. Of the ~2,400 errors coded by human raters in this corpus, LanguageTool recalled ~600 of these errors. The majority of the missed errors (~50%) were related to punctuation, with other errors including form, grammar, and word characteristics hovering around 10%. This result indicates a weakness of LanguageTool in that it misses many errors that human raters capture. In some cases, this may be a result of LanguageTool not having an extensive or accurate enough rule-based system to capture complex errors that may occur across phrases or clauses within sentences. It may also be the case that LanguageTool is not able to strongly capture the semantics of missing words or redundant words, which is a difficult task for rule-based software. However, understanding that comparison across studies are difficult to make, the recall rates reported by LanguageTool are similar to those reported for *Criterion*. For instance, Lavolette (2014) hand-coded 266 t-units (i.e., a dominant clause and its dependent clauses) and found that 206 of them contained errors, for which *Criterion* was able to identify errors in 111(45% recall). Importantly, as stated before, comparisons between studies are difficult. For example, the unit of analysis used by Lavolette (2014) was the t-unit and not individual errors, so it is difficult to know the recall accuracy for total errors in that analysis and, for the current analysis, errors were not calculated at the t-unit.

In terms of precision, it is important to discuss the errors that LanguageTool did flag that were not flagged by the initial human raters for the Writing-Pal corpus. There were 447 such errors of which 124 (28%) were rated as being both accurate and meaningful, 209 (47%) were rated as accurate, but not meaningful, and 114 (26%) were rated as neither accurate nor meaningful. For the 124 errors that the human raters missed that were judged to be both accurate and meaningful, the vast majority were spelling errors (n = 105, 85%). The remaining were grammar (n = 9), typographical (n = 1), whitespace (n = 4), and duplication (n = 1) errors. Thus, in a small number of the overall cases of errors flagged by LanguageTool, LanguageTool did find meaningful errors that had been missed by expert raters, but most of these errors were spelling errors. Of these, many were differences between single and compound words such as "where as" versus "whereas," "some day" versus "someday," and "every day" versus "everyday." An additional 209 errors were flagged by LanguageTool that were not flagged by the initial raters but were judged to be accurate, but not meaningful by the subsequent raters. The majority of these (n = 185) were labeled as English Quotes (uncategorized category) errors that were based on the use of straight quotes versus the suggested curly quotes (smart quotes). Thus, while LanguageTool was correct in noting that the authors had used straight quotes, the raters did not judge the use of straight quotes to be meaningful. Thus, while precision was low, much of the precision score was based on reporting errors that were accurate, but not meaningful.

Although LanguageTool did not recall the majority of errors in the Writing-Pal corpus, LanguageTool did demonstrate weak to strong correlations with expert judgments of grammar, mechanical, and syntax errors in that corpus. Specifically, the overall number of errors flagged by LanguageTool demonstrated a correlation of r = -0.514 with expert ratings. Individually, misspellings had the highest correlation with expert ratings followed by typographical and white space errors. All of LanguageTool's error counts demonstrated at least a small effect size with the expert ratings (r > .1, see

Table 10). Since the analytic scale did not distinguish between grammar, syntax, and mechanical errors, interpreting these results is difficult. It is possible that the human ratings most strongly reflect misspellings in the text and thus misspelling errors flagged by LanguageTool correlated the highest. It is also possible that grammar, syntax, and mechanical errors were equally important in influencing the expert rating and the correlations reflect greater accuracy on the part of LanguageTool in detecting misspelling and lower accuracies in detecting grammar and other types of errors. More precise and individual analytic ratings of grammar, syntax, and mechanics are needed to more strongly test relationships between human ratings and LanguageTool. Nevertheless, the findings indicate that while LanguageTool may miss many errors, the errors that it does capture seem to provide a generally accurate overall profile of a writer's structural and mechanical knowledge even though the strength of the overall relationship is strongly informed by spelling errors.

LanguageTool indices also showed significant correlations with essay quality, especially for L2 writers in the TOEFL corpus. For L2 writers, there was a strong, negative correlation between essay score and the total number of errors (r = -0.528). Medium and negative correlations were also reported for misspellings and grammatical errors. Negative and small relations were reported for all remaining errors except style errors which showed a positive correlation. Like the expert ratings of grammar, mechanical, and syntax errors reported in the Writing-Pal corpus, this finding supports the notion that LanguageTool provides a strong profile of writers' structural and mechanical abilities. This claim rests on the notion that less proficient L2 writers will produce more errors in their written texts, which is supported by the correlational analysis. Lastly, similar but weaker trends to the TOEFL corpus were reported for the Writing-Pal corpus in terms of essay quality. In the Writing-Pal corpus a number of LanguageTool indices showed small, negative relationships with essay quality including overall errors, misspellings, duplication errors, and grammar errors. The differences in the strength of the relationships supports previous findings that structural and mechanical errors are important predictors of L2 writing quality, but not L1 writing quality (Crossley et al., 2014).

7. Conclusion

While LanguageTool is freely accessible, programming knowledge is needed to access and retrieve textual information from it. Thus, we introduce GAMET, a tool that can automatically extract and count incidences of structural and mechanical errors in texts as calculated by LanguageTool using a GUI. The purpose of GAMET is to allow writing researchers without background knowledge in computer science or NLP to better explore relationships between writing metrics and structural and mechanical errors found in writing. GAMET is a desktop application that is freely available to researchers, works on major operating systems (Windows and Mac), and allows for batch processing of documents. A number of validation studies were conducted on features calculated by LanguageTool (and reported by GAMET). These studies investigated the accuracy and meaningfulness of the errors calculated by LanguageTool for both the Writing-Pal and the TOEFL corpora. In addition, the study examined how well the reported errors calculated for the Writing-Pal corpus match human error codings, and how strongly the errors reported by LanguageTool are associated with holistic judgement of L1 and L2 essay quality.

Overall, we find that LanguageTool seems to perform on-par with error correction tools designed to provide feedback to writers or automatically correct errors, noting, of course, that direct comparisons are not possible. While the precision of errors flagged by LanguageTool seem reliable, many errors are not recalled. This finding is offset by the notion that the errors flagged by LanguageTool do appear to provide a reliable profile of the overall number of errors in student writing as reported by the correlations between the incidence of errors and human ratings of writing quality and structural and mechanical errors. Therefore, LanguageTool and its instantiation in GAMET seems to provide an accurate profile of writers in terms of the structural and mechanical errors they produce. Future work should look to improve the capabilities of LanguageTool to better increase error recall, especially for punctuation and style errors. Style errors reported by LanguageTool seem to be unreliable and unrelated to human ratings of writing quality and should thus be assessed carefully before being used in subsequent analyses. Additionally, larger error coded samples are necessary to further test the recall abilities of LanguageTool and GAMET.

References

Attali, Y. (2008). *E-rater performance for TOEFL iBT independent essays*. Unpublished manuscript.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *Ai Magazine*, *25*(3), 27-27.

Crossley, S. A., Kyle, K., Allen, L, & McNamara, D. S. (2014). The Importance of Grammar and Mechanics in Writing Assessment and Instruction: Evidence from Data Mining. In Stamper, J., Pardos, Z., Mavrikis, M., & McLaren, B.M. (Eds.). *Proceedings of the 7th Educational Data Mining (EDM) Conference*. (pp. 300-303). Heidelberg, Berlin, Germany: Springer.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. Journal of Educational Psychology, 100, 907 – 919. https://doi.org/10.1037/a0012656

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. https://doi.org/10.1016/j.asw.2014.03.006

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25* (2), 155–185. https://doi.org/10.1177/0265532207086780

Ferris, D. R. (2011). *Treatment of error in second language student writing*. Ann Arbor: The University of Michigan Press.

Ferris, D., Eckstein, G., & DeHond, G. (2017). Self-directed language development: A study of first-year college writers. *Research in the Teaching of English*, *51*(4), 418.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication*, *32*(4), 365-387. https://doi.org/10.2307/356600

Graham, S. (1983). Effective spelling instruction. *Elementary School Journal, 83* (5), 560-567. https://doi.org/10.1086/461334

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of educational psychology*, *89*(1), 170. https://doi.org/10.1037//0022-0663.89.1.170

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445-476. https://doi.org/10.1037/0022-0663.99.3.445

Henderson, E. H., & Templeton, S. (1986). A developmental perspective of formal spelling instruction through alphabet, pattern, and meaning. *The Elementary School Journal, 86(3),* 304-316. https://doi.org/10.1086/461451

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of second language writing*, *12*(1), 17-29. https://doi.org/10.1016/s1060-3743(02)00124-8

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, *19*(2), 50-68.

Macdonald, N., Frase, L., Gingrich, P., & Keenan, S. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, *30*(1), 105-110. https://doi.org/10.1109/tcom.1982.1095380

Morris, D., Blanton, L., Blanton, W., & Perney, J. (1995). Spelling instruction and achievement in six classrooms. *Elementary School Journal, 96,* 145–162. https://doi.org/10.1086/461819

Myles, J. (2002). Second language writing and research: The writing process and error analysis in student texts. *TESL-EJ*, *6*(2), 1-20.

Naber, D. (2003). *A rule-based style and grammar checker*. GRIN Verlag.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1-14). https://doi.org/10.3115/v1/w14-1701

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. Educational Psychology, 37(1), 8-25. https://doi.org/10.1080/01443410.2015.1136407

Richards, J. C. (2015). Error analysis: Perspectives on second language acquisition. *Routledge*.

Rimrott, A., & Heift, T. (2008). Evaluating automatic detection of misspellings in German. *Language Learning & Technology*, *12*(3), 73-92.

Rozovskaya, A., Chang, K. W., Sammons, M., & Roth, D. (2013). The University of Illinois system in the CoNLL-2013 shared task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 13-19). https://doi.org/10.3115/v1/w14-1704

Santos, T. (1988). Professors' Reactions to the Academic Writing of Nonnative-Speaking Students. *TESOL Quarterly*, *22*(1), 69-90. https://doi.org/10.2307/3587062

Ward, S. A., & Seifert, R. (1990). The importance of mechanics in journalistic writing: A study of reporters and editors. *Journalism Quarterly*, *67*(1), 104-113. https://doi.org/10.1177/107769909006700116

Wei, L. (2015). Interlanguage: The abstract level in language acquisition. Edwin Mellen Press.

Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *Elementary School Journal, 99*(2), 101–128. https://doi.org/10.1086/461918

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL quarterly*, *16*(2), 195-209. https://doi.org/10.2307/3586792

Zhu, W. (2004). Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of second language Writing*, *13*(1), 29-48. https://doi.org/10.1016/s1060-3743(04)00007-4