Grammar Checkers Do Not Work

Les Perelman

Massachusetts Institute of Technology
Cambridge, Massachusetts

Daily I thank the powers that be for the computer spell checker. I never could spell decently. In grade school my work was always marked down for poor spelling. In undergraduate and graduate programs, I painstakingly reviewed papers with the *American Heritage Dictionary* to correct my numerous spelling mistakes. By the time I wrote my dissertation, I managed to cajole my then partner, now wife, to proofread it for spelling errors. (She is still collecting on that favor.)

All that changed in 1983 with WordPerfect's incorporation of a spell checker. My productivity as a scholar and teacher increased exponentially. Now when I type a spellchecked set of comments, I have no fear of embarrassing myself. Spell checkers also greatly influenced student writing. When Andrea Lunsford and Karen Lunsford's 2008 study reproduced the 1988 Robert Connors and Andrea Lunsford study of student writing errors, the greatest difference was in the major decline in the frequency of spelling mistakes. While spell checkers are often unable to identify homonyms such as *too* for *two*, overall they work well. Grammar checkers, however, do not work well.

The first grammar checkers, such as Writer's Workbench's grammar modules, began in the 1970s; MS Word and WordPerfect added grammar modules in the 1980s. By the late 1990s grammar checkers were mostly aimed at K-12 and post-secondary education—with products such as ETS's Criterion, Pearson Writer, and Measurement Incorporated's Project Essay Grade, along with stand-alone products such as Grammarly, WhiteSmoke (the grammar checker used by Pearson Education), and Ginger. We know spell checkers are usually accurate in detecting misspellings; that is, they are reliable. But are grammar checkers reliable?[1]

This question breaks down into several related ones:

- Does a grammar checker detect most, if not all grammatical errors?
- When it detects grammatical errors, does it correctly classify them in a manner that will allow writers to understand the errors and improve their writing?
- Does it classify some instances of perfectly grammatical prose as errors to produce false positives?

The answer to these questions is that grammar checkers are so unreliable that I can assert that they do not work.[1] At best, they detect around 50% of grammatical errors in a student text (Chodorow, Dickenson, Israel, and Tetreault; Gamon, Chodorow, Leacock, and Tetreault; Han, Chodorow, and Leacock). More troubling, because almost all grammar checkers use statistical modeling (more on that later), increases in the errors they identify will be accompanied by increases in false positives of perfectly grammatical prose being identified as an error (Gamon, Chodorow, Leacock, and Tetreault; Measurement Inc.). This phenomenon is most apparent when grammar checkers analyze an expert writer's prose. Using the online service WriteCheck, which employs the grammar checking modules from ETS's e-Rater,[2] I submitted 5,000 words (maximum allowed) from a favorite essay, "The Responsibility of Intellectuals" by Noam Chomsky. The ETS grammar checker found the following "errors" or "problems":

TABLE 1: WriteCheck Errors - Chomsky Article

Missing comma	9
Article error (missing or not needed)	15
Beginning sentence with **coordinating conjunction**	14
Spelling	4
Incorrect Preposition	5
Passive Voice	8
Sentence Fragment	2
Verb Form Error	1
Proofread. This part of the sentence contains a grammatical error or misspelled word that makes your meaning unclear.	2
Run-on sentence	1
Compound These two words should be written as one compound word.	1

Of the 62 problems identified in Chomsky's prose, only one could possibly be considered an error, a sentence fragment used for emphasis. All the other identified "errors" consisted of perfectly grammatical prose. The other sentence identified as a fragment was an independent clause with a subject and finite verb. I also ran a segment of 10,000 characters (maximum allowed) through WhiteSmoke. It identified 3 spelling errors, 32 grammar errors, and 32 problems in style.[3]

Grammar checkers often flag certain correct constructions as errors because those constructions are most often ones that computers can easily identify. Thus, although a sentence beginning with a coordinating conjunction has been accepted in almost all written English prose registers for at least 25 years, grammar checkers cling to the old rule because it is so easy for a computer to identify that "mistake." Once the algorithm has a list of the coordinating conjunctions, it simply tags any occurrence that begins a sentence. Similarly, most grammar checkers tag any introductory word or phrase, from *thus* to a prepositional phrase, that is not followed by a comma.

Articles and prepositions are difficult for machines to get right when they analyze the prose of expert writers, and they are difficult for English Language Learners. I ran a representative paper of 354 words from an advanced English Language Learner through seven grammar checkers: 1) MS Word; 2) ETS's e-Rater 2.0 in Criterion; 3) ETS's e-Rater 3.0 in WriteCheck; 4) Grammarly (free version); 5) Whitesmoke; 6) Ginger; 7) Virtual Writing Tutor; and 8) Language Tool. I identified 28 errors in the text, which I classified as major, middle, and minor errors. The 12 major errors consisted of incorrect verb forms or missing verbs; problems with subject-verb agreement; article misuse or omission; incorrect or missing preposition; and incorrect use of singular or plural noun form. I selected these errors because when they are read aloud, they are immediately apparent as errors to native speakers. The seven middle errors, still somewhat serious, included such problems as confusing shifts in verb tense and comma splices. The nine minor errors consisted almost entirely of missing commas, with one trivial usage problem.

Of the 12 major usage errors, one grammar checker identified only one error; two identified two errors; one identified three; one identified four; and two identified five errors. Three of the grammar checkers also each produced one false positive. These results largely replicate a more comprehensive study by Semire Dikli and Susan Bleyle, who compared error identification by two instructors and e-Rater 2.0 using 42 ELL papers. That analysis demonstrates that e-Rater is extremely inaccurate in identifying the types of major errors made by ELL, bilingual, and bidialectical students. The instructors coded 118 instances of missing or extra article; Criterion marked 76 instances, but 31 of those (40.8%) were either false positives or misidentified. One representative example of misidentification occurred when a student wrote the preposition *along* as two words *a long*, and Criterion marked it as an article error. The instructors coded 37 instances of the use of the wrong article; Criterion coded 17, but 15 (88.2%) of them, again, were either false positives or misidentified. The instructors coded 106 preposition errors, while Criterion identified only 19, with 5 of those (26.3%) being false positives or misidentified.

Grammar checkers don't work because neither of the two approaches being employed in them is reliable enough to be useful. The first approach is grammar-based. In the past 57 years, generative grammar has provided significant insights into language, especially syntax, morphology, and phonology. But the two other areas of linguistics—semantics, the meaning of words, and pragmatics, how language is used—still need major theoretical breakthroughs to be useful in applications such as grammar checkers. One main feature that governs the use of articles in English is whether a noun is countable or uncountable.[4] Although a class of English nouns is almost always countable, such as car, many other nouns are countable in some contexts and grammatical constructions but not in others:

1. Elizabeth saw a lamb.
2. Elizabeth won't eat lamb because she is a vegetarian.
3. Linguists seek knowledge of how language works.
4. Betty is developing a keen knowledge of fine wines.

Indeed, linguists now no longer classify nouns into the dichotomous categories of countable and uncountable, but have established various gradations of countability along a continuum (Allan; Pica).

Similarly, prepositions are appropriate in some contexts and not in others. Prepositions also serve multiple purposes. The preposition *by*, for example is used to indicate both the *instrumental case*, which indicates a noun is the instrument or means of accomplishing an action, and the *locative case*, which indicates a location. The major grammar checker currently employing a grammar-based approach is the one integrated into MS Word. The inherent flaws in employing such an approach with our limited linguistic knowledge, especially in the fields of semantics and pragmatics, can be easily demonstrated by writing in MS Word the following sentence with the grammar checker set to flag the passive voice:

> The car was parked by the side of the road.

MS Word will recommend the following revision:

> The side of the road parked the car.

Over time, MS Word has become more limited in what it flags. It no longer identifies article usage problems.

During the past 20 years, there has been a movement away from trying to build grammar-based grammar checkers to employing statistical analysis of huge corpora of data. This approach uses "big data" to predict which constructions are grammatical. A huge corpus of Standard English documents is fed into the machine, which performs regression analyses and other statistical processes to predict the probability that a construct in a new text is grammatical. The problem with such an approach is that it attempts to use an extremely large corpus of data to predict grammaticality for what is an infinite set of possible expressions in natural language. Even with immense computing power, this "big data" approach, like those used to predict winners at horse races,[5] stock market profit, or long-term weather forecasts, produces results that are not really useful. The sets of possible outcomes are simply too immense. In the case of grammar checkers, the imprecision of the statistical method translates as balancing the identification of all the errors present in the text against mistakenly tagging false positives, which will confuse students, especially bidialectical and bilingual students and English Language Learners. Overall, ETS's Criterion detects only about 40% of the errors in texts, while 10% of its reported errors are false positives. (Han, Chodorow, and Leacock). In identifying preposition use errors, Criterion only identifies about 25% of the errors present in texts, and about 20% of its tags on preposition use are false positives (Tetreault and Chodorow). In detecting article errors, Criterion correctly identifies only about 40% of the errors, while 10% of its reported errors are false positives (Han, Chodorow, and Leacock). The best results I have seen are those of a study by Measurement Inc., of its scoring engine, Project Essay Grade (PEG) in which both human readers and PEG marked almost 2000 sentences. PEG identified 52% of the errors and had only 1% false positives (Gamon, Chodorow, Leacock, and Tetreault). However, the specifics of this study, including descriptions of the specific writing task, the nature of texts, and the students who wrote them, are not reported.

Clearly, the inaccuracy of these statistical approaches is probably not helpful, and it is perhaps harmful to students. Although they often appear similar, grammar checkers are much more unreliable than spell checkers. Students can be easily deceived into thinking grammar checker corrections are comprehensive and reliable. They are not. Some grammar checkers warn that they may be inaccurate, but I have never seen one explicitly state something like “On average, 10% of the errors identified will not really be errors, and our product will identify only about 50% of errors in a student’s paper.”

For four years, Microsoft Research engineers worked on developing a grammar checker for English Language Learners based on statistical approaches (Chodorow, Gamon, and Tetreault) before discontinuing the project in 2011 (Gamon). Given that Microsoft has abandoned further work on grammar checkers, especially those using statistical approaches, and the other products on the market appear to be unreliable, why are grammar checkers still being used not only for classroom use, but also as a component of scoring engines for high-stakes tests?

The answer to this question is two-fold. First, with a few notable exceptions, such as the Dikli and Bleyle study, almost all research on the efficacy of grammar checkers has been done by researchers either employed by the organizations producing and selling the grammar checkers or by individuals associated with one or more of them. Many of the studies had no control group, or the control group consisted of students receiving no feedback at all (Chodorow, Gamon, and Tetreault). Second, that research community has redefined terms in an almost Orwellian fashion, which has made inaccurate grammar checkers seem precise. Researchers did away with the metric of accuracy, and substituted two measures, precision and recall (Chodorow, Dickinson, Israel, and Tetreault.) Deceptively, recall, the number of real errors detected by the system divided by the total number of real errors, is a transparent measure, except for its name, of accuracy. Precision, on the other hand, the number of correct errors identified by the system divided by the sum of the correct errors detected by the system and false positives, is simply a measure of how well a system avoids false positives, marking correct constructions as ungrammatical. In some cases, like this one, the researcher explains what the measures mean:

> "In detecting article errors, Criterion's precision is about 90% and its recall is about 40%. That means that when the system reports an error in a student's writing, the human annotator agrees about 90% of the time. However, Criterion detects only about 40% of the errors that the human marks (Han et al., 2006)." (Chodorow, Gamon and Tetreault p. 427)

However, in many instances the terms are used without explanation, and an administrator hearing that the precision of the system is 90% might be impressed while wondering what *recall* means.

Although the abilities of grammar checkers already look unimpressive, additional questions need to be raised and researched.

- Grammar checkers are used as a major component of Automated Essay Scoring (AES) machines used in high stakes testing from K-12 and the TOEFL through graduate admissions tests such as the Graduate Record Examination (GRE) and the Graduate Management Admissions Test (GMAT). Given the unreliability of grammar checkers, should AES machines continue to be used in high stakes testing?

- What are the effects of grammar checkers on student writing, especially on ELL, bilingual, and bidialectical students? Specifically, what are the effects of randomly (at least to anyone but the computer) identifying some errors but not others? Clearly, random marking of papers differs from the individualized approach of writing center consultants helping students understand, within the context of their own style and a specific genre, why an error is a mistake and how to avoid it in future writings.[6]

- Do anomalies inherent in the statistical techniques of grammar checkers privilege ELLs from some language groups and discriminate against others through skewed false negatives and false positives?

- How do grammar checker limitations affect arguments,

> such as that by Paul Deane, supporting a role for computers in assessing specific components of the writing construct such as grammar and mechanics?

Finally, when administrators want you to use automated tools for instruction in grammar and mechanics, ask for pieces of their best prose and a favorite Op-Ed and run them through ETS's e-Rater by spending $10-$20 at <en.writecheck.com>. The results should end the conversation.

[1] In addition to studying automated grammar checkers, I have been a critic of computer evaluation of writing, both in the classroom and in high-stakes testing, such as the essay portion of the SAT. See the *New York Times* article by Michael Winerip and *The Chronicle of Higher Education* article by Steve Kolowich.

[2] ETS denied me access to the new Criterion version unless I allowed them advanced review of any presentation or publication and the option to require removal of any reference to ETS or their products from the presentation or publication. I obtained e-Rater 3.0 access by buying a limited subscription to WriteCheck.

[3]Screen shots of the output from WriteCheck's analysis are available at <lesperelman.com/writing-assessment-robo-grading/parts-noam-chomskys-essay-grammar-checked/>.

[4] Interestingly as I write this essay in MS Word, a green squiggly line has appeared under *governs* because the parser cannot recognize that the subject of the sentence is singular.

[5] See Michael Nunamaker (2001). <web.archive.org/web/20011109073203/http://nationalturf.com/nunamaker/>.

[6] I assume such assistance is offered after rhetorical concerns have been discussed or when such discussion is not needed.

◆ ◆ ◆ ◆ ◆

Allan, Keith. " Nouns and Countability." *Language*. 56.3 (1980) 541–567. Print.

Chodorow, Martin, Markus Dickinson, Ross Israel, and Joel Tetreault. "Problems in Evaluating Grammatical Error Detection Systems." *Proceedings of COLING: Technical Papers*. Mumbai: COLING, (2012) 611-628. Web. 31 Jan. 2016.

Chodorow, Martin, Michael Gamon, and Joel Tetreault. "The Utility of Article and Preposition Error Correction Systems for English Language Learners: Feedback and Assessment." *Language Testing* 27.3 (2010). 419-436. Print.

Chomsky, Noam. "The Responsibility of Intellectuals" *The New York Review of Books*. 23 Feb. 1967. Web. 31 Jan. 2016.

Connors, Robert J., and Andrea A. Lunsford. "Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research." *College Composition and Communication* 39.4 (1988): 395-409. Print.

Deane, Paul. "On the Relation Between Automate d Essay Scoring and Modern Views of the Writing Construct." *Assessing Writing* 18.1(2013): 7-24. Print.

Dikli, Semire, and Susan Bleyle. "Automated Essay Scoring Feedback for Second Language Writers: How Does It Compare to Instructor Feedback?." *Assessing Writing* 22 (2014): 1-17. Print.

Gamon, Michael. "ESL Assistant Discontinued." *Microsoft Research*. 28 Apr. 2011. Web.

Gamon, Michael, Martin Chodorow, Claudia Leacock, and Joel Tetreault. "Grammatical Error Detection in Automatic Essay Scoring and Feedback." *Handbook of Automated Essay Evaluation: Current Applications and New*

Directions. Ed. Mark D. Shermis and Jill Burstein. New York: Routledge (2013):251-266. Print.

Han, Na-Rae, Martin Chodorow, and Claudia Leacock. “Detecting Errors on English Article Usage by Non-Native Speakers.” *Natural Language Engineering* 12.2 (2006): 115-129. Print.

Herrington, Anne and Charles Moran. “Writing to a Machine is Not Writing at All.” *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White.* Ed. Norbert Elliot and Les Perelman. New York: Hampton. (2012) 425-438. Print.

Kolowich, Steve. “Writing Instructor Skeptical of Automated Grading, Pits Machine vs. Machine.” *Chronicle of Higher Education*. Web. 28 Apr., 2014, sec. Technology. 30 Apr., 2014.

Lunsford, Andrea A., and Karen J. Lunsford. “’Mistakes are a Fact of Life’: A National Comparative Study.” *College Composition and Communication* 59.4 (2008): 781-806. Print.

Measurement Inc. PEG Writer FAQ: 8. “Why Does PEG Seem to Ignore Some Grammar ‘Trouble Spots’ Identified by Microsoft Word (Or Other Programs)?” Web. 7 Feb. 2016.

Pica, Teresa. “The Article in American English: What the Textbooks Don’t Tell Us.” *Sociolinguistics and Language Acquisition*. Ed. Nessa Wolfson and Elliot Judd. Rowley, MA: Newbury. 1983. 222–233. Print.

Tetreault, Joel R., and Martin Chodorow. “The Ups and Downs of Preposition Error Detection in ESL Writing.” *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. (2008). Manchester, UK: Association for Computational Linguistics. 865-872. Print.

Winerip, Michael. “Facing a Robo-Grader? Just Keep Obfuscating Mellifluously.” *New York Times*. Web. 22 Apr. 2012, sec. Education. 14 Feb., 2016.

Looking for more good reading about writing center work? There’s the blog, “Connecting Writing Centers Across Borders” (a global connection for all writing centers). Post your news on Twitter and Facebook pages, and use WcORD to search for links to web resources on writing centers:

WLN blog:	www.wlnjournal.org/blog/
WLN Twitter:	twitter.com/WLNjournal
WLN Facebook:	www.facebook.com/wlnjournal
WcORD:	wlnjournal.org/wcord.php