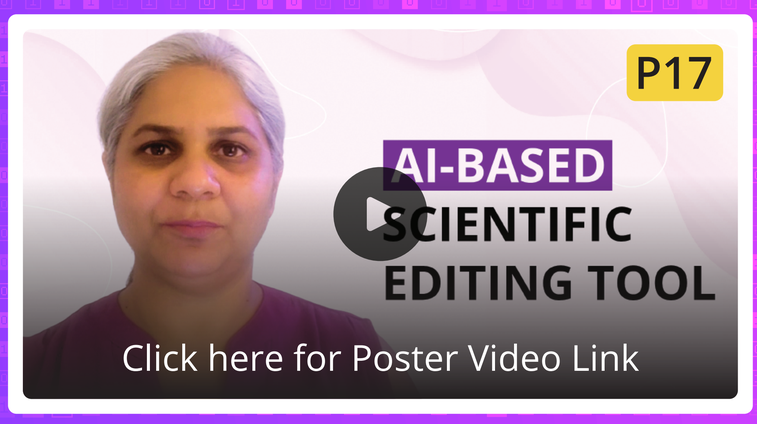
AI-based Scientific Editing Tool (SET) aids in improving the linguistic quality of scientific documents and accelerates publication

Trinka D'Cunha[a] and Pinki Rajeev[a]
[a]Enago Life Sciences-Crimson Interactive Pvt. Ltd., Mumbai, India



OBJECTIVE

- Researchers/authors are often in a rush to get their research published. To expedite scientific paper submissions, reduction of time and effort spent in ensuring linguistic accuracy is critical.
- Moreover, research output is rapidly increasing globally,[1] making it a difficult task for journal editors to screen new submissions efficiently.
- With the aim to automate essential editorial checks, we designed an AI-based Scientific Editing Tool (SET)—a tool that guides authors/publishers in writing and screening scientific documents for linguistic accuracy, thereby adding efficiency to the publication process.

RESEARCH DESIGN AND METHODS

- We designed an Artificial Intelligence-based Scientific Editing Tool (SET) that uses neural networks trained on academic writing and linguistic rules. The neural networks learned from ~15 million scientific sentences and rules that were carefully crafted by linguists and copyeditors.
- SET is trained on and built for scientific editing with a focus on academic manuscripts.
- SET is designed to perform 27 broad language checks representing more than 3000 error types across grammar, syntax, vocabulary, spelling, UK/US style, punctuation, and formal/unbiased language categories and provides a language quality score (Q-score) **(Table 1)**.
- We compared the editing quality of 50 manuscripts (~3000 words/manuscript) from the medicine and biosciences domains; these manuscripts were edited by copyeditors and SET.

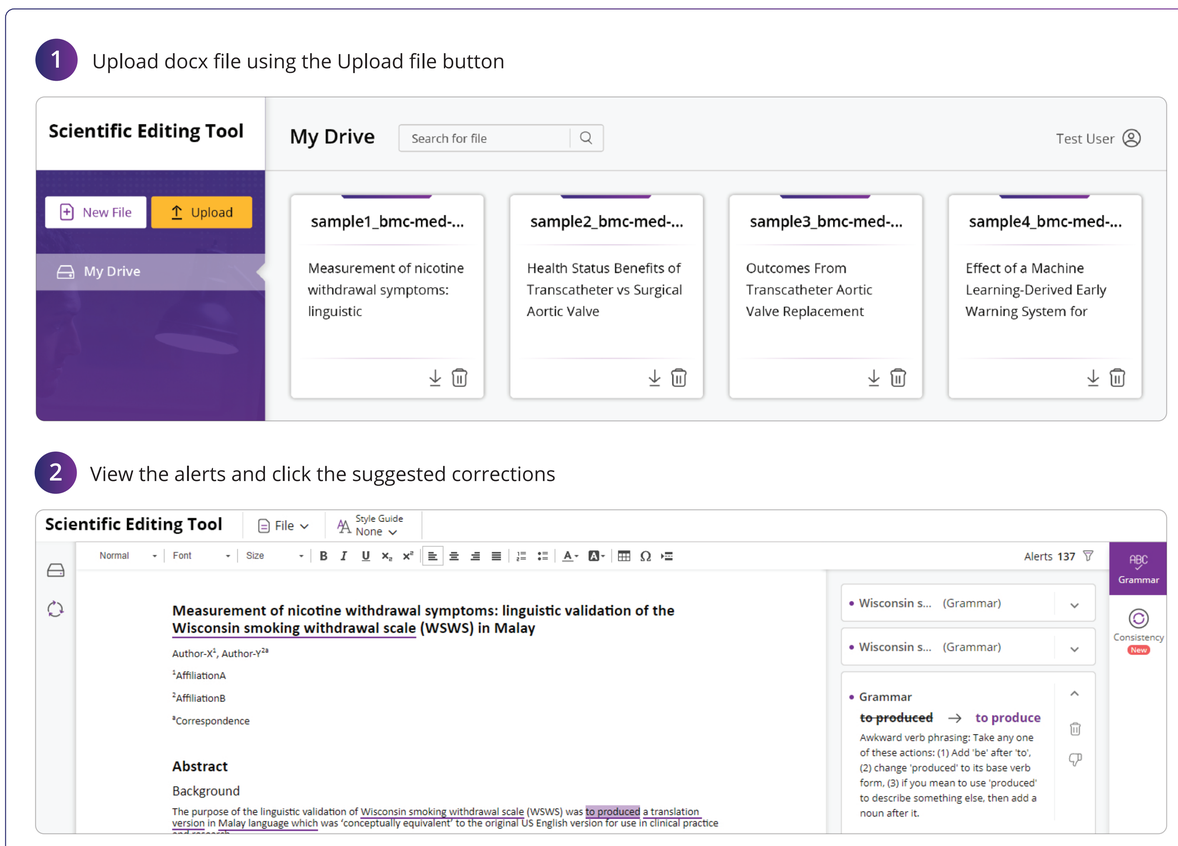
Table 1. Language Checks Performed by SET

Category	Description
Grammar	Articles, determiners, noun number, pronouns, subject-verb agreement, verb inflections, word forms, adjective/adverbs, tense, prepositions, and conjunctions
Syntax & Vocabulary	Run-on sentences, fragments, dangling modifiers, simple parallelism, word order, word/phrase choice, confusables, unbiased/inclusive language, redundancy, conciseness, sentence structure, and idioms
Punctuation	Commas, hyphens/dashes, semicolons, colons, and apostrophes
General writing style	US–UK style conventions, spelling, spacing, capitalization, word choice, academic register, and tone
Scientific style guide	AMA 11, APA 7, and AGU 2017
Consistency check	Consistency in use of dashes, number style, spacing, Greek letters, and symbols

RESULTS

- With a single click, the user-friendly, interactive SET analyzed the manuscript and generated editorial corrections in approximately five seconds **(Figure 1)**.

Figure 1. SET Interface



- SET performed more than 3000 language checks, including scientific expressions, and provided the user with clickable language correction advice **(Table 2)**.

Table 2. Linguistic Corrections Suggested by SET

Category	Correction Example
Grammar: Articles, subject–verb agreement, noun number	Therefore, full information about aromatic side chains of **the** microenvironment ~~were~~ **was** found, such as differences in vibrational ~~modes~~ **mode** intensity and intensity ratio alteration of Fermi resonance doublets.
Syntax & Vocabulary: Word/Phrase choice	Here, we ~~experience~~ **describe** a surviving case from a BO-LVFWR occurring during a PCI for an AMI.
Punctuation: Colons & Semicolons	There are roughly two kinds of removal ~~methods;~~ **methods:** endoscopic resection and surgical resection. *(The semicolon has been changed to the colon.)*
General writing style	**This** is to ensure the optimal transfer of the original message and measuring what is intended to be measured. *(When using the pronoun 'this' without a noun, make sure that the noun or thought it is referring to can be easily understood by the reader. Else, add a suitable noun after 'this' to make your meaning clear.)* **Fifteen**~~15~~ subjects were selected to undergo fat volume measurement via abdominal CT. *(Avoid using numerals at the beginning of a sentence.)*
Scientific style guide	By the 11th day, she had a **fever** of 105°F and fluid in the right pleural cavity. *(Use 'elevated temperature' instead of 'fever' and specify the temperature in parenthesis. [AMA 11])*
Consistency check	This **patient-reported outcome** measure combines the strength of... (pg. 2) Development of **patient reported outcome** measures such as WSWS for... (pg. 7) *(Variations in hyphenated words found.)*

- The Q-scores of the unedited sample manuscripts (N = 50) were <60% (poor); 60%–79% (moderate); and >80% (good). Post SET editing, the Q-score improved by 36%, 22%, and 7%, respectively.
- There was a 35% (±15%) direct and 23% (±13%) indirect overlap between editors and SET editorial suggestions. A direct edit overlap occurs when the editor and the tool make the same change, whereas an indirect overlap occurs when the editor and the tool correct an error in different ways.

  Example:
  - The sentence "To get the information about Raman spectrum structure, it was essential to classify **marker bands which represents** protein interaction and structure" has a subject–verb error (bands which represents).
  - The **editor** changed "which represents' to 'representing'; The **tool** changed it to 'that represent'. Both edits correct the error, but differently.
- We found that SET made 25% (±11%) additional corrections which editors overlooked. These grammatical overlooks pertained to articles, noun number, punctuation, subject–verb agreement and prepositions.
- Editors, however, corrected and enhanced the manuscript and made various in-depth text revisions to improve the clarity and presentation of the document that were not fixed by SET.

Corrections for clarity/presentation suggested by Editors but missed by SET

Example 1: Information re-ordering

Unedited: Implementation of AI technologies into endocytoscopy has been eagerly explored by researchers at Showa University in Japan.

Edited: *Researchers at Showa University in Japan have eagerly explored the implementation of* AI technologies into endocytoscopy.

Example 2: Readability

Unedited: However, the progress of research on automated prediction of cancer invasion is in a very early phase compared with that of automated polyp detection and characterization because of the lack of number of invasive cancers which is far less than that of colorectal polyps.

Edited: However, the progress of research on automated *invasive cancer prediction* is in a very early phase compared with that of automated polyp detection and characterization because *the number of* invasive cancers *is far lower* than that of colorectal polyps.

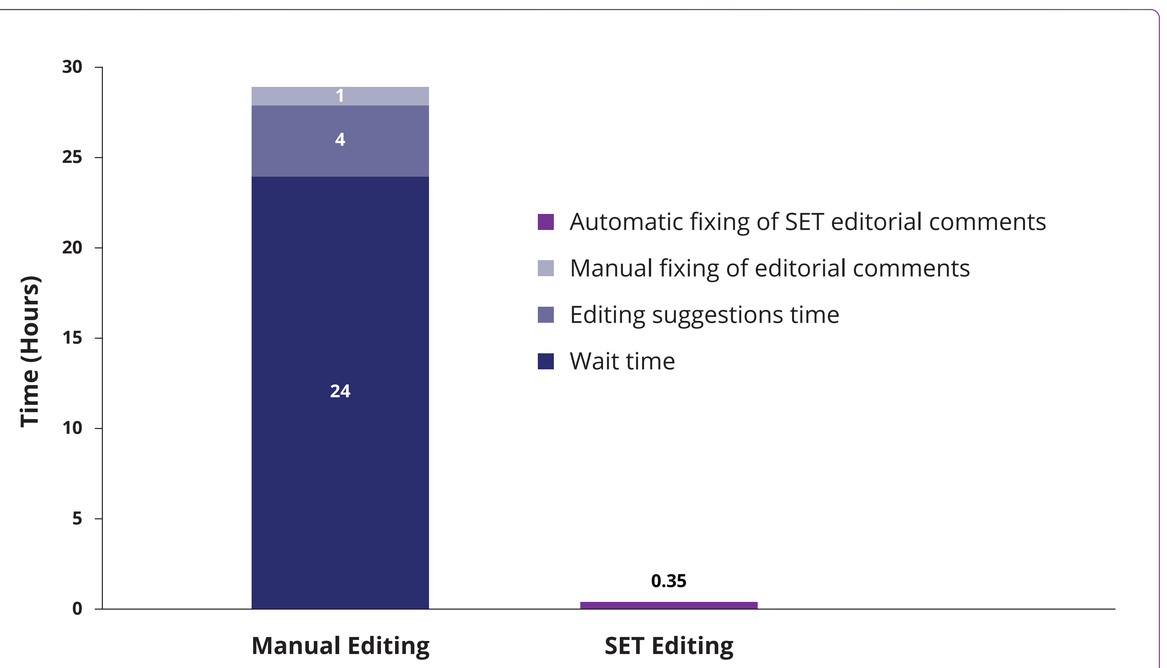
Example 3: Clarity & Concision

Unedited: Acne scar that is atrophic is found around 85%–90%.

Edited: The occurrence of atrophic acne scars is found *to be* around 85%–90%.

- Over manual editing, SET reduced editing time by ~5 hours and turn-around-time by ~24 hours **(Figure 2)**.

Figure 2. Average Editing Time Taken (~3000 words): Manual vs. SET



CONCLUSIONS

- SET improves the quality of scientific language and helps reduce the time required for editing a document.
- With SET handling basic linguistic errors, editors can focus on fixing deeper issues and can help authors/publishers accelerate their editing operations.

LIMITATIONS AND FUTURE SCOPE

- SET provides significant benefits to authors, publishers, medical writers, and editorial operations; however, there is scope for improvement. SET covers errors caught while proofreading (e.g., articles, verb agreement, noun number, articles, commas etc.) while language issues requiring logical reasoning or real-world knowledge are not handled. This function is best performed by a human editor.
- SET provides corrections using the sentence as context; hence, error coverage in verb tense between sentences and pronoun agreement is less. Such corrections require document knowledge, which is a feature for future development.

Reference

1. Johnson R, Wakinson A & Mabe, M. (2018). The STM Report. An overview of scientific and scholarly publishing 1968–2018. Celebrating the 50th anniversary of STM. International Association of Scientific, Technical and Medical Publishers. Retrieved from https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf (Accessed: January 7, 2021)

Acknowledgements

Authors would like to thank the Crimson AI team for developing SET; Sachin Rane for sharing valuable insights; Amit Garg for his critical review; the design team at Enago Life Sciences for their design support.