

Alfred Russel Wallace Notes 42: How Accurate Are the Transcriptions Presented at *The Alfred Russel Wallace Page Website?*

Charles H. Smith, PhD,^a August 2026

^a Professor Emeritus, Western Kentucky University, Bowling Green, KY. Email: charles.smith@wku.edu

Summary: A look is taken at the level of accuracy displayed by the transcriptions of Wallace writings offered at the *Alfred Russel Wallace Page* website, as determined by a ChatGPT analysis. *Key words:* Alfred Russel Wallace, transcriptions, ChatGPT, editing

I take my responsibility as a Wallace editor/transcriber quite seriously, so I thought it might be relevant to provide some new evidence as to the worthiness of the transcriptions of Wallace material found at my *Alfred Russel Wallace Page* website (see <https://people.wku.edu/charles.smith/wallace/writings.htm>). This is not just me crowing; users have a right to know what levels of accuracy have been enforced there.

An initial version of the Wallace service offering only a couple of dozen transcriptions was opened on the website of UCAL Berkeley's Museum of Vertebrate Paleontology in 1998. In 2000 I moved it to a server hosted by my employer (Western Kentucky University), and immediately began to implement a rigorous expansion. One goal for it was to eventually provide transcripts of all of Wallace's published writings. As at that time some eight hundred such writings were known to exist and I had no outside funding, I envisioned a long-term effort. (Since then I have identified a further more than three hundred previously unrecognized titles, making the job even more daunting.) Soon after the migration in 2000, I wrote the following words atop the 'Full-Text of Wallace's Published Writings' page at the site:

...The items have been very diligently proofread to correct transcription errors, though I cannot of course guarantee that absolutely none remain. Still – As I additionally have been able to correct a fair number of previously existing typos, the renditions offered are probably on the whole little or no more afflicted with errors than were the originals themselves.

These are fighting words, but, before attempting to defend them, I should say a few words about my process of transcription.

Up to the present, just about all of the transcripts – and *all* of the ones that are not books – have been presented in an html format initiated through word-processing. The shorter items have usually been typed in *de novo*, whereas many of the longer ones were the products of an initial photo scan, followed by an OCR operation yielding a plaintext file. Surprisingly, perhaps, the photo scanning approach was hardly any faster than hand-entering a typed script up front, for two reasons.

First, many of the originals had to be scanned in one page at a time, and then the OCR'd files reformatted line by line so they would wrap properly as a word processing/html entity.

Second, most of the real time spent on each item actually involved the proofreading process: that is, with checking, visually and word by word, each word processed/plaintext version against the source material. Each 'page' of outputted material, speaking on the average, took some ten to twenty minutes to proofread, and this was always done at least twice (usually by myself, and a second person) to ensure accuracy of transcription. Spellcheck aids were of course also employed to identify glaring errors, but these were not too heavily counted on because they couldn't flag omissions, duplications, and most full-word or full-phrase mistakes.

Beyond this, the images to be scanned were often or usually of such poor clarity that the OCR strategy could not be profitably applied at all. In my possession were copies of hundreds, or thousands, of pages of Wallace writings that had been produced by early-generation photocopiers (or worse yet, microfilm reader printers), and were barely legible to the naked eye, much less an automated algorithm. I was in many instances also at the mercy of whatever quality of item initial copy that interlibrary loan services were able to beg from their sources.

Nowadays all of this could be done somewhat faster, but only somewhat. I haven't a formal count, but Wallace publications appeared in something like three hundred *different* venues, many of which are very obscure. This is a prime reason why Prof. John van Wyhe was happy to make use of my transcriptions for most of the Wallace text items on his service *Wallace Online* when I offered them to him: had he not taken me up on the offer, his site would have required several years longer to produce, even with his substantial funding aids to speed up the effort.

So, now to return to the main matter. Just recently (July 2026) I decided to undertake a probable final review of the transcriptions on my site, taking advantage of free services offered by the A.I. provider ChatGPT. When directed to do so, the ChatGPT algorithm can perform content searches on texts made available to it – specifically, it can identify possible or probable internal errors directly, or compare a subject text to some other version of the same. I quickly decided it would take too long to create electronic images of all my source material (or find such online), so instead settled for providing the captured text of my transcriptions, and asking the algorithm to search for anything that possibly looked like a transcription error on my part. Making the request only takes a couple of minutes for each of the 'S' numbered Wallace writings at my site – though it of course takes much longer to check up on all the problematic items the algorithm identifies through its analysis: there may be anywhere from ten to more than fifty such items identified, and each must then be visually (manually) investigated (that is, through comparisons of each queried online html transcription passage to a matching printed copy one). This process usually takes between about ten to thirty minutes per Wallace publication.

Note that merely finding original electronic copies of Wallace's writings for direct comparisons doesn't help much anyway, as the OCR'd text derived from these is often

messy enough to further complicate the process of comparison.

At this point I have just completed the first 112 of these analyses/comparisons (total time required: about thirty hours); these involve 'S' numbers S1 through S100, originating in the years 1843 to 1864 (Wallace's early years through to the third year after his return to England from the East). So, what did I find?

I am pleased to report that ChatGPT has proclaimed my transcriptions as being "extraordinarily faithful to the originals." More specifically, the algorithm found a total of only 32 errors in the 112 items, or on the average about one for every 3½ publications. Looked at in another way, since only 21 of the 112 displayed any errors at all, about 81% of the latter were determined to be error-free. The character-by-character mistake rate is a bit harder to calculate, but it was on the order of about 0.01 or 0.02 percent, or around one or two in ten thousand.

This does not mean, of course, that the transcriptions can be claimed as literally "error free"; still, along the way I've also corrected a small handful of what were obvious printing errors (possibly, in some instances, stemming from occasions on which the typesetters could not read Wallace's handwriting, or they simply goofed), so on the balance I believe I can still reasonably claim that the transcriptions "are probably on the whole little or no more afflicted with errors than were the originals themselves." This also assumes, of course, that ChatGPT is as good as it seems to be at detecting possible errors – but I have little reason to believe otherwise, as the algorithm also rates the likelihood that the particular contentious items it identifies represent true transcription errors as opposed to period grammatical or spelling oddities, and it has almost always been on target.

The actual errors discovered, and corrections made thereto, were in the following 'S' numbers:

anacaris > aracaris (S6)
same > some (S7)
athletics > athletic; hated > hatred; course > coarse (S13a)
compromising > comprising (S15)
tree > three; sage > sago (S18)
haring > having (S25)
keep > kept; abject > object (S26)
of > or (S29)
delicate; > delicate, (S37)
Ptiddae > Pittidae (S38)
pp. 169 > p. 169; 105 > 105° (S41)
ever > even (S47)
I to 100 > 1 to 100 (S65)
wings > wing (S69)
Trans. > Trans., (S71)
alhes > allies; suhapical > subapical; 1 7/10 > 1 7/10 (S72)
require > required (S82)
1 9/10 > 1 9/10; alls > alis; niarginatis > marginatis; pirima > prima (S85)
. . > . xviii.; Picidae > Picideae (S88)
eternal > external (S98)

It should be fairly easy to guess which of these were manual transcription mistakes, and which were overlooked OCR errors. In the S98 one, for example, it is very unlikely that an OCR operation would simply have missed this one letter entirely; instead, it more probably would have *mis-identified* a letter. All of these corrections have already been applied to the online texts.

Thirty hours is a reasonably small total, but as there are now more than eleven hundred in all to look at, it will probably be the better part of a year before the task is completed.

Acknowledgments

My thanks to reviewers Rosemary Meszaros, George Beccaloni, and Brian Coutts.

Earlier Titles in this Series*

Note 41. "Violet Wallace's 'Natural History for Infants' " (July 2026).

Note 40. "A Comprehensive Bibliography of Books and Booklets about Alfred Russel Wallace" (April 2026).

Note 39. "What About Wallace's 'Agnosticism'...?" (March 2026).

Note 38. "Alfred Russel Wallace's 'Revolt of Democracy': A Critical Analysis of Early 20th Century Labor Theory and Its Contemporary Relevance" (October 2025).

Note 37. "What About Wallace on Vaccination...?" (July 2025).

Note 36. "Wallace, Evolution and Providence" (April 2025).

Note 35. "Natural Selection by Evolution" (March 2025).

Note 34. "Did Alfred Russel Wallace's Chief Assistant from His Expedition to the Malay Archipelago Really Name Himself Ali Wallace?" (March 2025).

Note 33. "The Story in Evolution" (December 2024).

Note 32. "Wallace on the Tigers of Singapore" (November 2024).

Note 31. "Wallace's Personal Copy of 'Life of Alexander von Humboldt'" (August 2024).

Note 30. "Wallace a Theist? Part 1." (April 2024).

Note 29. "Alien Entertainment: Are We Being Monitored by Alien Beings?" (January 2024).

Note 28. "Wallace's 'Change of Mind', Revisited" (November 2023).

Note 27. "When Wallace Broke with Darwin" (October 2023).

Note 26. "Confessions of a 'Wallace Enthusiast'" (May 2023).

Note 25. "Wallace and the 'Physical Environment'" (April 2023).

Note 24. "Wallace at 200: Potential Subjects for Student Theses" (December 2022).

Note 23. "The Jersey Devil, and Friends" (December 2022).

- Note 22. "Stumbling Blocks to an Understanding of Wallace's Worldview" (August 2022).
- Note 21. "Wallace and the Doorway to the Universe" (May 2022).
- Note 20. "Did Darwin and Wallace 'Coauthor' the 1858 Communication on Natural Selection?" (April 2022).
- Note 19. "Social Evolution's Useful Idiots" (February 2022).
- Note 18. "Wallace on the Balance of Nature" (August 2021).
- Note 17. "More on the South Asian Connection" (April 2021).
- Note 16. "The Flexible Wallace" (April 2021).
- Note 15. "Wallace's Many 'Hats': What Should We Call Him?" (January 2021).
- Note 14. "Background for Wallace's 1845 Kington Essay" (November 2020).
- Note 13. "Wallace on Prayer" (November 2020).
- Note 12. "How Good Was Wallace's Memory?" (November 2020).
- Note 11. "Wallace, Darwin, Education, and the Class Question" (October 2020).
- Note 10. "The Impact of A. R. Wallace's Sarawak Law Paper Resurrected" (April 2020).
- Note 9. "The South Asian Connection" (2019).
- Note 8. "Wallace's Earliest Exposures to the Writings of Alexander von Humboldt" (October 2018).
- Note 7. "Wallace, Bates, and John Plant: The Leicester Connection" (October 2017).
- Note 6. "More on the Mailing Date of the Ternate Essay to Darwin" (April 2015).
- Note 5. "Just How Well Known Was Wallace in His Own Time?" (April 2014).
- Note 4. "Contributions to 'The Garden', 1875-1912" (October 2011).
- Note 3. "Two Early Publications" (October 2011).
- Note 2. "The Spelling 'Russel', and Wallace's Date of Birth" (October 2010).
- Note 1. "Authorship of Two Early Works" (April 2010).

*Available through ResearchGate, TopScholar, or on request from Charles H. Smith. Beginning with Note 10, Alfred Russel Wallace Notes is a refereed, irregularly published, note series edited by Charles H. Smith, available open access through the Western Kentucky University electronic repository.